

An Investigation of a Human in the Loop Approach to Object Recognition

Orod Razeghi, BSc.

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

July 2015

UNIVERSITY OF NOTTINGHAM

Abstract

Faculty of Science
School of Computer Science

Doctor of Philosophy

by Orod Razeghi

For several decades researchers around the globe have been avidly investigating practical solutions to the enduring problem of understanding visual content within an image. One might think of the quest as an effort to emulate human visual system. Despite all the endeavours, the simplest of visual tasks to us humans, such as optical segmentation of objects, remain a significant challenge for machines. In a few occasions where a computer's processing power is adequate to accomplish the task, the issue of public alienation towards autonomous solutions to critical applications remains unresolved.

The principal purpose of this thesis is to propose innovative computer vision, machine learning, and pattern recognition techniques that exploit abstract knowledge of human beings in practical models using facile yet effective methodologies. High-level information provided by users in the decision making loop of such interactive systems enhances the efficacy of vision algorithms, whilst simultaneously machines reduce users' labour by filtering results and completing mundane tasks on their behalf.

In this thesis, we initially draw a vivid picture of interactive approaches to vision tasks prior to scrutinising relevant aspects of human in the loop methodologies and highlighting their current shortcomings in object recognition applications. Our survey of literature unveils that the difficulty in harnessing users' abstract knowledge is amongst major complications of human in the loop algorithms. We therefore propose two novel methodologies to capture and model such high-level sources of information. One solution builds innovative textual descriptors that are compatible with discriminative classifiers. The other is based on the random naive Bayes algorithm and is suitable for generative classification frameworks.

We further investigate the infamous problem of fusing images' low-level and users' high-level information sources. Our next contribution is therefore a novel random forest based human in the loop framework that efficiently fuses visual features of images with user provided information for fast predictions and a superior classification performance. User abstract knowledge in this method is harnessed in shape of user's answers to perceptual questions about images. In contrast to generative Bayesian frameworks, this is a direct discriminative approach that enables information source fusion in the preliminary stages of the prediction process.

We subsequently reveal inventive generative frameworks that model each source of information individually and determine the most effective for the purpose of class label prediction. We propose two innovative and intelligent human in the loop fusion algorithms. Our first algorithm is a modified naive Bayes greedy technique, while our second solution is based on a feedforward neural network. Through experiments on a variety of datasets, we show that our novel intelligent fusion methods of information source selection outperform their competitors in tasks of fine-grained visual categorisation.

We additionally present methodologies to reduce unnecessary human involvement in mundane tasks by only focusing on cases where their invaluable abstract knowledge is of utter importance. Our proposed algorithm is based on information theory and recent image annotation techniques. It determines the most efficient sequence of information to obtain from humans involved in the decision making loop, in order to minimise their unnecessary engagement in routine tasks. This approach allows them to be concerned with more abstract functions instead. Our experimental results reveal faster achievement of peak performance in contrast to alternative random ranking systems.

Our final major contribution in this thesis is a novel remedy for the curse of dimensionality in pattern recognition problems. It is theoretically based on mutual information and Fano's inequality. Our approach separates the most discriminative descriptors and has the capability to enhance the accuracy of classification algorithms. The process of selecting a subset of relevant features is vital for designing robust human in the loop vision models. Our selection techniques eliminate redundant and irrelevant visual and textual features, and therefore its influence on improvement of various human in the loop algorithms proves to be fundamental in our experiments.

List of Publications

1. Razeghi, O.; Qiu, G.; Williams, H. and Thomas, K. (2012), Skin Lesion Image Recognition with Computer Vision and Human in the Loop, in ‘Medical Image Understanding and Analysis’, pp. 167–172.
2. Razeghi, O.; Qiu, G.; Williams, H. and Thomas, K. (2012), Computer Aided Skin Lesion Diagnosis with Humans in the Loop ‘Machine Learning in Medical Imaging’, Springer Berlin Heidelberg, pp. 266–274.
3. Razeghi, O.; Zhang, Q. and Qiu, G. (2013), Interactive Skin Condition Recognition, in ‘IEEE International Conference on Multimedia and Expo’, pp. 1–6.
4. Razeghi, O.; Fu, H. and Qiu, G. (2013), Building Skin Condition Recogniser using Crowd-sourced High Level Knowledge, in ‘Medical Image Understanding and Analysis’, pp. 225–230.
5. Razeghi, O.; and Qiu, G. (2014), 2309 Skin Conditions and Crowd-sourced High-level Knowledge Dataset for Building a Computer Aided Diagnosis System, in ‘IEEE International Symposium on Biomedical Imaging’, pp. 61–64.
6. Razeghi, O.; and Qiu, G. (2014), Object Recognition with Human in the Loop Intelligent Framework, for ‘Journal of Pattern Recognition’, [under review].
7. Razeghi, O.; and Qiu, G. (2014), Discriminative Dimension Reduction based on Mutual Information, for ‘Journal of Pattern Recognition’, [under review].

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this study. I am thankful for their aspiring guidance, invaluable constructive criticism and friendly advice. I am sincerely grateful to them for sharing their truthful and illuminating views on all aspects of this thesis.

I owe my deepest gratitude to Professor Guoping Qiu, whose encouragement, supervision and support from the preliminary stages to the concluding levels enabled me to develop and finalise this project. He has been a tremendous mentor for me. I would like to thank him for allowing me to grow as a young scientist.

I am sincerely grateful to Professor Natasha Alechina, who has been a fantastic guide and a friendly face to look upon even at hardship since my undergraduate years. I would also like to thank Professor Tony Pridmore for serving as my second supervisor.

I have accumulated a debt of gratitude to many colleagues and friends including Dr. Ghasemi, Dr. Moradi, Dr. Fu, Dr. Hirbod, Bozhi, Qian, Jie, Venon, Reza, Siamak, Pejman, and others to whom I cannot express my appreciation in words.

I would like to express my sincerest gratitude to my Dad, to my two sisters Ghazal and Atregol, and to the rest of my family whose support has been enormous. A few sentences here will never be enough to describe their extremely invaluable worth to me.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this study.

Thank you all.

Orod Razeghi

Contents

Abstract	i
List of Publications	iii
Acknowledgements	iv
List of Figures	ix
List of Tables	x
Abbreviations	xii
Symbols	xiii
1 Introduction	1
1.1 Prologue	1
1.2 Motivation	3
1.3 Aims and Objectives	4
1.4 Structure and Contributions	6
2 Literature Review	9
2.1 Features in Computer Vision	10
2.1.1 Feature Extraction	10
2.1.2 Features of Higher Level	11
2.1.3 Feature Representation	12
2.1.4 Feature Combination	13
2.1.5 Feature Selection	14
2.2 Classification in Computer Vision	15
2.2.1 Bayesian Classifiers	15
2.2.2 Support Vector Machine	16
2.2.3 Multiple Kernel Learning	19
2.2.4 Artificial Neural Networks	20
2.2.5 Ensemble Classification	22
2.2.5.1 Random Forest	22
2.2.5.2 Random Naive Bayesian	24

2.3	Understanding Visual Content at Pixel Level	25
2.3.1	Object Detection	25
2.3.2	Image Segmentation	27
2.4	Understanding Visual Content at Image Level	27
2.4.1	Visual Content Classification	28
2.4.2	Content Based Image Retrieval	30
2.4.3	Automatic Image Annotation	30
2.5	Understanding Visual Content by Human in the Loop	32
2.5.1	Visual Content Classification	32
2.5.2	Image Segmentation	34
2.5.3	Semi-automatic Image Annotation	35
2.5.4	Medical Applications of Human in the Loop	36
2.5.5	Information Source Fusion in Human in the Loop Applications	37
2.5.6	Current Shortcomings in Human in the Loop Approaches	39
2.6	Summary	39
3	Discriminative Object Recognition with Humans in the Loop	41
3.1	Problem Formulation	43
3.2	Representation of Low-level Image Information	44
3.3	Representation of High-level User Information	45
3.4	Human in the Loop Random Forest Classifier	46
3.4.1	Random Forest Construction	46
3.4.2	Classification in the Random Forest	47
3.5	Experiments and Results	47
3.5.1	Derm90 and Derm706 Skin Conditions Dataset	47
3.5.1.1	Experiment Setup	49
3.5.1.2	Baseline Results	50
3.5.1.3	Random Forest Results	51
3.5.2	Derm2309 Skin Conditions Dataset	53
3.5.2.1	Experiment Setup	55
3.5.2.2	Baseline Results	56
3.5.2.3	Random Forest Results	57
3.5.3	MIAS Mammographic Dataset	58
3.5.3.1	Experiment Setup	60
3.5.3.2	Baseline Results	60
3.5.3.3	Random Forest Results	60
3.5.4	Caltech-UCSD Birds 200 Dataset	61
3.5.4.1	Experiment Setup	63
3.5.4.2	Baseline Results	63
3.5.4.3	Random Forest Results	63
3.5.5	Ground Photograph Habitat Dataset	64
3.5.5.1	Experiment Setup	66
3.5.5.2	Baseline Results	67
3.5.5.3	Random Forest Results	67
3.6	Conclusion	68
4	Generative Object Recognition with Humans in the Loop	70

4.1	Problem Formulation	71
4.2	Intelligent Information Source Fusion	73
4.2.1	A Modified Naive Bayes Information Fusion Algorithm	73
4.2.2	Neural Network Fusion Algorithm	75
4.3	Classifiers in Fusion Frameworks	76
4.3.1	A Generative Model for High-level User Information	77
4.3.1.1	Presentation of High-level Information	77
4.3.1.2	Modelling User Answers	78
4.3.1.3	Ensemble of Random Naive Bayes Classifiers	79
4.3.2	A Discriminative Classifier for Image Low-Level Information	79
4.4	Experiments and Results	80
4.4.1	Derm2309 Skin Conditions Dataset	81
4.4.1.1	Visual Results	81
4.4.1.2	Textual Results	81
4.4.1.3	Combination Results	81
4.4.2	MIAS Mammographic Dataset	82
4.4.2.1	Visual Results	82
4.4.2.2	Textual Results	83
4.4.2.3	Combination Results	83
4.4.3	Caltech-UCSD Birds 200 Dataset	84
4.4.3.1	Visual Results	84
4.4.3.2	Textual Results	84
4.4.3.3	Combination Results	85
4.4.4	Ground Photograph Habitat Dataset	86
4.4.4.1	Visual Results	86
4.4.4.2	Textual Results	87
4.4.4.3	Combination Results	87
4.5	Conclusion	88
5	Ranking Order of Human Information	89
5.1	Problem Formulation	90
5.2	Automatic Answer Prediction	92
5.2.1	Construction of a Random Forest for Tag Prediction	92
5.2.2	Tag Prediction by the Random Forest Algorithm	93
5.3	Experiments and Results	95
5.3.1	Derm90 and Derm706 Skin Conditions Datasets	95
5.3.1.1	Order of User Answers	95
5.3.1.2	Frequency of User Answers	96
5.3.2	Derm2309 Skin Conditions Dataset	97
5.3.2.1	Automatic Answers Accuracy	98
5.3.2.2	Questions Ranking Effect	98
5.3.3	MIAS Mammographic Dataset	99
5.3.3.1	Automatic Answers Accuracy	99
5.3.3.2	Questions Ranking Effect	100
5.3.4	Caltech-UCSD Birds 200 Dataset	100
5.3.4.1	Automatic Answers Accuracy	101
5.3.4.2	Questions Ranking Effect	101

5.4	Conclusion	102
6	Discriminative Subspace Selection based on Mutual Information	103
6.1	Problem Formulation	104
6.2	Mutual Information Subspace	105
6.2.1	Examples of Common Subspace Methods	108
6.2.2	Useful Properties	109
6.3	Experiments and Results	110
6.3.1	Synthetic Data	110
6.3.2	Derm2309 Skin Conditions Dataset	112
6.3.3	MIAS Mammographic Dataset	113
6.3.4	MSRC 21-class Dataset	113
6.3.5	Oxford Flower Recognition Dataset	114
6.3.6	Pascal VOC2007 Challenge Dataset	115
6.3.7	UCI Machine Learning Repository Datasets	117
6.3.8	Yale Face Recognition Dataset	119
6.3.9	Interpretation of Results	120
6.4	Conclusion	121
7	Concluding Remarks	122
7.1	Main Contributions	122
7.2	Limitations and Future Work	124
7.3	Epilogue	125
A	Derm2309 Skin Conditions Dataset	127
A.1	Data Format	127
A.2	Method to Read	127
A.3	Evaluation Criteria	128
A.4	Download Address	128
	Bibliography	129

List of Figures

1.1	Al-Jazari's Automaton	2
1.2	Human in the Loop Algorithm	4
2.1	Support Vectors	17
3.1	Skin Recognition Tool with Human in the Loop	42
3.2	Graphical User Interface (GUI) in Interactive Medical Experiments	46
3.3	Textual Results of Derm706 Dataset	52
3.4	Example Images from Derm2309 Dataset	53
3.5	Amazon Mechanical Turk Interface	55
3.6	Random Forest Accuracy on Derm2309 Dataset	58
3.7	Example Images from MIAS Dataset	59
3.8	Example Images from Caltech-UCSD Birds 200 Dataset	62
3.9	Example Images from Ground Photograph Habitat Dataset	65
4.1	Fusion Frameworks at Input and Output Levels	72
4.2	Neural Network Layout	76
5.1	Logical Flow of Selecting Suitable Questions	91
5.2	Semantic Neighbours	94
5.3	Answers Frequency for Derm706 Dataset	97
5.4	Question Ranking Effect on Derm2309 Dataset	99
5.5	Question Ranking Effect on MIAS Dataset	100
5.6	Question Ranking Effect on CUB-200 Dataset	102
6.1	Mutual Information Subspace on Toy Dataset	111

List of Tables

2.1	Segmentation Algorithms	28
3.1	User Answers Certainties	45
3.2	Individual Visual Features' Accuracies on Pilot Dermatology Datasets . .	48
3.3	Derm90 Dataset Questions	49
3.4	Derm706 Dataset Questions	50
3.5	Derm90 Dataset Classification Accuracies	51
3.6	Derm706 Dataset Classification Accuracies	51
3.7	Derm90 Dataset Individual Class Accuracies	51
3.8	Derm706 Dataset Individual Class Accuracies	52
3.9	Full List of Derm2309 Skin Conditions	54
3.10	Derm2309 Dataset Questions	56
3.11	Derm2309 Dataset Classification Accuracies	57
3.12	Derm2309 Dataset Individual Class Accuracies	58
3.13	MIAS Dataset Questions	59
3.14	MIAS Dataset Classification Accuracies	61
3.15	Caltech-UCSD Birds 200 Dataset Questions	62
3.16	Caltech-UCSD Birds 200 Dataset Classification Accuracies	64
3.17	Ground Photograph Habitat Dataset Questions	66
3.18	Ground Photograph Habitat Dataset Classification Accuracies	68
4.1	Classifiers Settings in Human in the Loop Frameworks	77
4.2	User Answers Certainties	78
4.3	Mean Accuracy of Classification Algorithms on Derm2309 Dataset	81
4.4	Mean Accuracy of Fusion Algorithms on Derm2309 Dataset	82
4.5	Mean Accuracy of Classification Algorithms on MIAS Dataset	83
4.6	Mean Accuracy of Fusion Algorithms on MIAS Dataset	84
4.7	Mean Accuracy of Classification Algorithms on CUB-200 Dataset	85
4.8	Mean Accuracy of Fusion Algorithms on CUB-200 Dataset	86
4.9	Mean Accuracy of Classification Algorithms on Habitat Dataset	87
4.10	Mean Accuracy of Fusion Algorithms on Habitat Dataset	88
5.1	Order of Derm90 Questions Asked by the Ranking Algorithm	96
5.2	Order of Derm706 Questions Asked by the Ranking Algorithm	97
6.1	Mean Accuracies in Percentage on Derm2309 Dataset	112
6.2	Mean Accuracies in Percentage on MIAS Dataset	114
6.3	Mean Accuracies in Percentage on MSRC 21-class Dataset	115
6.4	Mean Accuracies in Percentage on Oxford Flower Dataset	116

6.5	Mean Accuracies in Percentage on Pascal VOC2007 Dataset	117
6.6	Mean Accuracies in Percentage on UCI-Sonar Dataset	118
6.7	Mean Accuracies in Percentage on UCI-MFeat Dataset	119
6.8	Mean Accuracies in Percentage on Yale Face Dataset	120
6.9	Difference in Bases returned by MI and the Conventional Methods	121

Abbreviations

AMT	A mazon M echanical T urk
CBIR	C ontent B ased I mage R etrieval
CPAM	C oloured P attern A ppearance M odel
DCT	D iscrete C osine T ransform
GB	G eometric B lur
HOG	H istogram of O riented G radients
IG	I nformation G ain
KPCA	K ernel P rincipal C omponent A nalysis
LDA	L inear D iscriminant A nalysis
MI	M utual I nformation
MKL	M ultiple K ernel L earning
MLE	M aximum L ikelihood E stimation
NB	N aive B ayesian
NN	N eural N etwork
PCA	P rincipal C omponent A nalysis
PHOG	P yramid H istogram of O riented G radients
PHOW	P yramid H istogram O f visual W ords
RF	R andom F orest
RNB	R andom N aive B ayes
RP	R andom P rojection
SIFT	S cale I nvariant F eature T ransform
SSIM	S tructure S imilarity I ndex M easure
SVM	S upport V ector M achine

Symbols

c	Class Label
F	Feature Representation
G	Training Samples
H	Entropy
I	Information
S	User Textual Information
U	Universal Source of Information
x	Image Visual Information
θ	Threshold Value(s)

*“Of knowledge naught remained I did not know,
Of secrets, scarcely any, high or low;
All day and night for three score and twelve years,
I pondered, just to learn that naught I know.”*

Omar Khayyam, 11th Century
Persian Polymath, Philosopher, and Poet

Dedicated To My Mother

Chapter 1

Introduction

1.1 Prologue

Mechanical reasoning has been contemplated by philosophers and mathematicians since antiquity. Thinking machines and artificial entities first began to appear in ancient Greek myths, such as Talos of Crete, and the bronze robot of Hephaestus. It is widely acknowledged that by the middle ages artificial beings had been created by polymaths and scholars like Jabir ibn Hayyan, Judah Loew and Paracelsus. History is filled with stories of humanoid automatons built by intellectuals like Yan Shi, Hero of Alexandria, and Al-Jazari who is renowned for writing the book of “Knowledge of Ingenious Mechanical Devices” in 1206. The hand washing automaton illustrated in figure 1.1 for instance is amongst the hundred devices he carefully described in his book. By the 19th and the early 20th centuries, artificial beings had become a common feature in fiction as in Mary Shelley’s acclaimed *Frankenstein* or Rossum’s *Universal Robots* by the Czech writer Karel Capek.

Pamela McCorduck, the author of “*Machines Who Think*” [1], argues that all of these efforts are examples of an ancient urge. A desire “to forge the gods”, as she describes it. Stories of these creatures and their fates debate many of the similar hopes, fears and ethical concerns that are presented by modern artificial intelligence today.

An intriguing branch of modern artificial intelligence is machine learning which is concerned with the construction and study of systems that can learn from data. In machine learning, pattern recognition centres around the identification of patterns and regularities in data. All these domains have evolved substantially from their roots in artificial intelligence, engineering and statistics but yet they have become increasingly similar by integrating developments and ideas from each other. As a scientific discipline



FIGURE 1.1: A hand washing automaton of Al-Jazari

that is highly correlated with these subject matters, computer vision is concerned with the theory behind artificial systems that extract useful information from images as the source of data. This thesis addresses the problem of semantic image understanding. Its ultimate objective is to reveal the semantic meaning behind the pixels of an image.

In the rest of this chapter, we highlight the motivations behind our proposed work, followed by challenges associated with our consciously defined aims and objectives. We eventually conclude this chapter by listing our novel contributions to the resolution of the challenges described.

1.2 Motivation

Outstanding strides have been made in the field of digital imaging over the past years. Digital images are now ubiquitous. From digital SLR¹s and megapixel camera phones to scanners and video surveillance systems, all these devices have made their notable mark on our modern lives. The Internet has clearly been the catalyst in fostering the growth of digital imaging. This rapid pace of growth has necessitated an urge to store, retrieve, and understand visual content intelligently. Intelligent digital imaging has already established its outstanding value in a variety of fields ranging from education to medicine [2].

The outlasting problem of understanding the visual content of an image has been vigorously scrutinised by the computer vision community. Researchers have been extensively engaged in designing innovative methodologies for capturing, processing, analysing, and understanding image data from the real world to engineer useful information in form of practical decisions. Albeit research on computer vision tasks dates back to the earliest days of computing and despite the huge growth of interest it has seen in recent years, humans still consistently outperform state-of-the-art computer vision algorithms at most tasks both in terms of accuracy and efficiency.

Attempts have eagerly focused on closing the so-called *sensory* and *semantic* gaps [2]. In simple words, while the former describes the gap between real world objects and their virtual representations in a computational space derived from recording of those objects, the latter portrays the gap between extraction of information from some visual data by computer systems and user interpretation of the same visual data in a given situation. An ideal system should be able to narrow these gaps by capturing accurate representations of a user's semantic requests in order to be competent enough at providing desirable outputs.

Despite all the efforts in the last few decades to overcome the aforementioned problems of understanding visual content and its corresponding complications, it is naive to believe that the-state-of-the-art techniques have matured adequately to serve our modern needs for most real world applications. Simply, this is due to the fact that accuracies of such technologies are not always solid. Furthermore, it is imperative to remind ourselves that members of the public do not trust machines to decide independently for execution of critical tasks [3], such as automatic diagnosis in medical applications. Therefore, we think that there should be further systematic psychological research carried out on the impact of technology on humans and their reactions to such intelligent entities.

¹Single-lens Reflex

The difficulties described thus far have gradually opened a new point of view in the community. It is known that the most advanced technologies still have a long way to become reliable, and it is clear that the issue of trust and public alienation towards autonomous technologies will not disappear overnight. Hence among various approaches, a number of researchers have recently shown interest in developing more pragmatic solutions to the problem by introducing the notion of “Human in the Loop”. The high-level knowledge of humans corrects any mistakes made by the algorithms. In return, interactive algorithms decrease human labour in many mundane tasks. Having a human in the decision making loop will also assure any doubts remaining in minds regarding authentication and accuracy of a solution.

An illustrative example [4] in the paradigm of human in the loop is a computational algorithm that finds a solution interactively. The solution may or may not satisfy the user’s expectation. If the answer is affirmative, a solution has been found. If not, the user will interact with the system to provide feedback that contains both human high-level knowledge about the problem, as well as their intentions. This feedback will be harnessed by the algorithm as more accurate constraint conditions or as stronger priors to refine the proposed solution. This loop of human in the computational process can be iterated until a satisfactory result is achieved. Figure 1.2 depicts this interactive approach in an imaging setting.

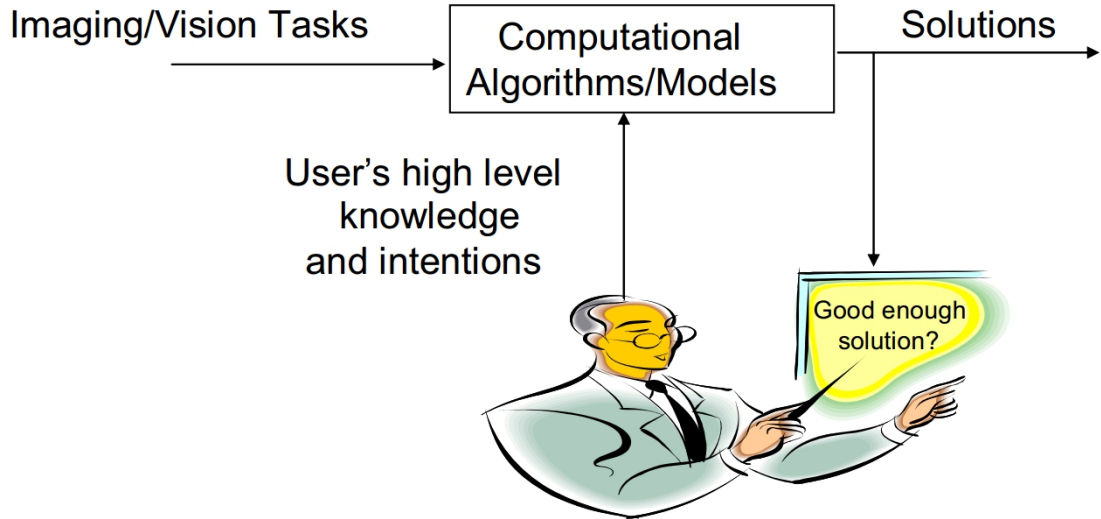


FIGURE 1.2: An illustration of interactive approaches to imaging and vision

1.3 Aims and Objectives

The semantic image understanding problem can be tackled from a number of different perspectives: i) image classification or annotation deals with determining whether an

image contains a certain object, ii) object detection attempts to locate the presence of an object within an image, and iii) image segmentation aims to assign a semantic label to every pixel in an image. These problems may seem different on surface but they all share intrinsic mutualities.

The difficulty of these tasks lies primarily in the huge volume of variability in real world images. The pixel intensities that generate any given digital image are broadly dependent on a large number of complex factors such as pose, appearance, shape, illumination etc. Hence, it can be considered that extracting information about the real world from the patterns of light that fall onto a camera's sensor is ill-posed. We think that the issue of high intra-class variation, and low inter-class variation demands accurate inferences with internal models that effectively combine sensory evidence with prior knowledge about the properties of real world objects. In this thesis, we ultimately aim to answer the following questions that are fundamental in the realisation of a visually intelligent automaton, which is capable of accurate inferences subject to a degree of assistance from humans in its decision making loop:

- 1) A practical feature extraction method is the first building block of such accurate algorithmic inferences. An ideal feature descriptor should be invariant to change amongst various entities within the same semantic class, and should be discriminative in differentiating between plausible semantic classes. It is usually vital to combine a number of different feature descriptors to obtain a comprehensive understanding of visual content within an image. This propagates to the further questions about effectiveness of each feature combination, removal of redundant features, and how to efficiently combine these feature descriptors.
- 2) Apart from low-level information attainable from visual feature descriptors, there exists a void in formalising methodologies for capturing valuable high-level knowledge from users in human in the loop settings. Their prior knowledge about properties of real world can substantially enhance the accuracy of a learning method by overcoming the semantic gap problem. Thus, precise scrutiny is obliged to develop methodologies that harvest such beneficial knowledge.
- 3) Besides the aggravations caused by the information representation dilemma described previously, selecting appropriate learning algorithms competent at recognising patterns is a constitutional necessity. Accurate parametrisation of such algorithms is another obstacle that we aim to precisely tackle.
- 4) We believe that utilisation of human abstract knowledge in the preceding algorithm has to be executed in an arrangement that reduces the burden of its users. Humans

should be concerned with abstract tasks rather than tedious and mundane assignments. Achieving a balance in this scenario is another angle that needs to be explored.

5) Finally, decisive fusion of low-level visual information and high-level abstract knowledge is the last significant component of an intelligent algorithm that is capable of accurate inference. The accomplishment of effective fusion demands certain questions to be answered and a number of difficult issues to be rectified. The most critical complication is the evaluation of predictions based on each source of information available.

We believe that it is becoming progressively common to frame computer vision problems as that of inference in probabilistic models. This allows for more convenient reasoning about higher level image concepts, a principled way to express uncertainty, and separation of model design and inference. The aim of this thesis is to follow these described schools of thoughts to tackle the technical challenges mentioned. We aim to develop platforms where modern technologies help with the problem of understanding visual content, and to implement user friendly interfaces that efficiently incorporate high-level human knowledge.

1.4 Structure and Contributions

In this thesis, we will formally define what is meant by human in the loop algorithms, illustrate how effective models can be learned from available sources of data, and describe how these different sources of data, comprising low-level image information and high-level human knowledge, can be combined to extract valuable information. This thesis makes the following novel contributions in the field of interactive understanding of visual content, and thence the remainder of this document is structured as below:

Chapter 2: presents an overview of fundamental constituents in computer vision and pattern recognition problems, current methodologies for content understanding at pixel and image level, and finally followed by an analytical comparison of existing work in the area of interactive imaging and object recognition with human in the loop.

Chapter 3: discloses a novel “Random Forest” based human in the loop framework that efficiently fuses visual features of images with user provided information for fast predictions and a superior classification performance. User abstract knowledge in this method is harnessed in shape of user answers to perceptual questions. These responses are used to build textual features compatible with random forest classifiers. In contrast to the generative Bayesian frameworks in chapter 4, this is a direct discriminative

approach that enables information source fusion in the preliminary stages of the prediction process. The work of this chapter has been published as follows:

1. Razeghi, O.; Zhang, Q. and Qiu, G. (2013), Interactive Skin Condition Recognition, in ‘IEEE International Conference on Multimedia and Expo’, pp. 1–6.
2. Razeghi, O.; Fu, H. and Qiu, G. (2013), Building Skin Condition Recogniser using Crowd-sourced High Level Knowledge, in ‘Medical Image Understanding and Analysis’, pp. 225–230.

Chapter 4: introduces a “Random Naive Bayes” model of capturing human high-level information that is compatible with the human in the loop Bayesian frameworks, in addition to innovative “Human in the Loop Fusion Frameworks” that intelligently select the most effective source of information suitable for making predictions. Through experiments on a variety of human in the loop datasets, we demonstrate the advantages of our “Random Naive Bayes” model in comparison to state-of-the-art methods both in terms of accuracy and efficiency. We also show that our novel intelligent methods of “Information Source Selection” outperform their competitors in tasks of fine-grained visual categorisation. The work presented in this chapter has been published as follows:

1. Razeghi, O.; Qiu, G.; Williams, H. and Thomas, K. (2012), Skin Lesion Image Recognition with Computer Vision and Human in the Loop, in ‘Medical Image Understanding and Analysis’, pp. 167–172.
2. Razeghi, O.; Qiu, G.; Williams, H. and Thomas, K. (2012), Computer Aided Skin Lesion Diagnosis with Humans in the Loop ‘Machine Learning in Medical Imaging’, Springer Berlin Heidelberg, pp. 266–274.
3. Razeghi, O.; and Qiu, G. (2014), Object Recognition with Human in the Loop Intelligent Framework, for ‘Journal of Pattern Recognition’, [under review].

Chapter 5: reveals a novel method to reduce unnecessary human intervention in decision making procedures. Our proposed algorithm determines the most “Efficient Sequence of Information” to obtain from humans involved in the decision making loop, in order to minimise their unnecessary engagement in mundane tasks. This approach allows them to be concerned with more abstract functions instead. The work presented in this chapter has been published in:

1. Razeghi, O.; Fu, H. and Qiu, G. (2013), Building Skin Condition Recogniser using Crowd-sourced High Level Knowledge, in ‘Medical Image Understanding and Analysis’, pp. 225–230.

Chapter 6: demonstrates a novel remedy for the curse of dimensionality in pattern recognition problems that is based on “Mutual information and Fano’s Inequality” methods. Our approach separates the most discriminative descriptors and has the capability to enhance the accuracy of many classification algorithms. The process of selecting a subset of relevant features is vital to designing robust human in the loop vision models. Our selection techniques eliminate redundant or irrelevant visual and textual features. The work presented in this chapter is illustrated in:

1. Razeghi, O.; and Qiu, G. (2014), Discriminative Dimension Reduction based on Mutual Information, for ‘Journal of Pattern Recognition’, [under review].

Chapter 7: summarises the results presented in the thesis, outlines directions for future work and concludes with a discussion.

Appendix A: exhibits a unique medical dataset containing 2309 images from 44 different skin conditions, which is suitable for human in the loop approaches. We believe that this dataset will be useful in facilitating the development of computer-aided medical diagnostic techniques. We have made the extracted low-level visual descriptors of images in the dataset, and the crowd-sourced high-level knowledge of users publicly available in the following paper:

1. Razeghi, O.; and Qiu, G. (2014), 2309 Skin Conditions and Crowd-sourced High-level Knowledge Dataset for Building a Computer Aided Diagnosis System, in ‘IEEE International Symposium on Biomedical Imaging’, pp. 61–64.

Chapter 2

Literature Review

The current computer vision literature has been profoundly engaged in exploring a range of topics that can be categorised into pixel-level semantic image understanding, and image-level visual content understanding respectively. Acquiring high-dimensional data, detection of certain objects within an image, semantic segmentation, and building more efficient mathematical models are prime examples of the former category, whilst interpreting high-level knowledge, recognition of different classes of data, image retrieval and image annotation are usually considered to be from the latter category. These topics address the core problems in the context of interpreting semantic meaning from images. In spite of their apparent differences, they all share a number of analogies. The issue of representing images in confined mathematical models, also known as features, is one common module. Classification is a key instance of another shared module amongst most of the topics previously mentioned.

As this thesis concentrates on solving the problem of understanding visual content based on a human in the loop approach, we will examine recent innovations in the field of semantic understanding with a focus on relevant complications of involving humans in the decision making loop. In the rest of this chapter, we will first review relevant methods in feature extraction as the first building blocks of any semantic understanding system. Subsequently, we will look into object categorisation and classification algorithms of such systems. We then review object detection and semantic segmentation as instances of understanding visual content at pixel-level. Afterwards, we will survey image-level semantic understanding methodologies. We will be discussing automatic image annotation, content based image retrieval, and finally we will summarise recent human in the loop developments in the literature. We will conclude this chapter with a short summary.

2.1 Features in Computer Vision

A feature is defined to acquire visual properties of an image, either locally for small patches of pixels or globally for the entire image. Features are perhaps the most significant concept in computer vision. They are used to signify information that is relevant for solving a computational task in a certain application. Features eliminate the need to deal with pixels in computer vision tasks by abstracting the complexity of data within an image. Classifiers are usually applied directly to extracted features of an image.

There are various methods proposed in the literature that represent an image by detecting points of interest and extracting meaningful descriptors from them. Feature extraction algorithms based on their processing primitives are usually divided into three basic categories: pixel-level, regional, and image-level. The following summarises a number of popular feature extraction methodologies and widely used representation techniques in the computer vision community.

2.1.1 Feature Extraction

Each point in an image is commonly represented by a value from the colour channels, such as *RGB*, *CMYK*, or *HSV*. These colour values are accompanied by a location (x, y) that corresponds to the position of that point in the image. However, a pixel is usually placed in a larger spatial context when it comes to feature extraction at pixel-level. A patch that centres on a given pixel is usually exploited to derive descriptors. These extracted descriptors are thenceforth perceived to belong to that central pixel.

The authors in [5] for instance propose an extraction method that covers a large area centred on a pixel. Randomly cropped rectangles within this large space are used to extract features. Assembled two-tuples that contain both extracted features and location of the rectangles are then utilised as the feature of that particular pixel. This leads to an exponential possibilities of features for any pixel within an image.

In contrast to pixel-level features, descriptors can be extracted at regional-level based on the following aspects:

- **Colour:** Colour Name [6], Colour SIFT [7], etc.
- **Texture:** Texton Histogram [8], etc.
- **Shape:** Histogram of Oriented Gradients (HOG) [9], etc.

- **Geometry:** Super Pixels [10], etc.
- **Appearance:** Scale Invariant Feature Transform (SIFT) [11], etc.

The authors in [12] intriguingly propose a kernel view of different regional features. Their kernel descriptor technique directly turns pixel attributes, such as gradient, colour, and local binary patterns into concise regional features. Kernel descriptors are claimed to be straightforward to design and therefore can turn any type of pixel attribute into regional-level features.

Contrastingly, CPAM [13] is introduced as a method of representing achromatic and chromatic image signals independently. An opponent colour representation, human vision theories, and modern signal processing technologies are combined to develop a computationally efficient visual appearance model for coloured image patterns. The opponent colour vision models achromatic and chromatic signals differently. The former should be provided with a higher bandwidth and the latter's signal can cope with a lower bandwidth. The normalisation of these signals has the effect of removing lighting condition to some extent. Consequently, the CPAM model demonstrates some illumination invariant properties. A compact representation of these patterns is achieved by vector quantisation [14], which is a well-developed statistical technique in the field of modern digital signal processing. CPAM in summary captures statistical representation of achromatic and chromatic spatial image patterns and uses their distribution to characterise visual content of an image.

At image-level, GIST [15] descriptors are amongst competent candidates. A typical GIST feature contains Gabor orientation histograms calculated over patches in a regular grid. GIST features are tuned by default parameters such as: 3 colour planes, 4 by 4 cells, and 3 scales with different orientations. These default parameters will produce a standard vector of 960 dimension. The colour histogram features, and fisher vectors [16] are also considered to be in this category.

2.1.2 Features of Higher Level

Most features described in the previous section are known as “bottom-up” features due to their intrinsic nature of computation. However, it is desirable to mention that there has been comparatively recent interest in using the outputs of classifiers as a new type of higher level features. Authors in [17] for instance examine the outputs of various object detectors as new features for image classification problems. An extended version of this idea can be found in [18], where the output of numerous individual action detectors is utilised as new features for action recognition.

A common characteristic of the aforementioned feature methods is that they are all in general handcrafted and comparatively simple. They typically follow a procedure of:

1. dense sampling of local image patches,
2. describing patches by means of visual descriptors such as SIFT [11],
3. encoding descriptors into a high-dimensional representation,
4. and finally pooling over the entire image.

Recently, these handcrafted approaches have been substantially outperformed by the introduction of the latest generation [19] of Convolutional Neural Networks (CNN) [20, 21] to the computer vision field. These networks have a considerably more sophisticated structure than standard representations. They are comprised of several layers of non-linear feature extractors, and are therefore said to be *deep*. This is in contrast to classical representation methods that are known to be referred to as *shallow*. Whilst the structure of these deep methods is handcrafted, they contain a considerable number of parameters learnt directly from data.

We have come to believe that considering the current rapid growth of interest in feature representation techniques based on deep learning [22–24] and their superiority of performance in many application settings [19], it is not inconceivable to observe a decline in popularity of other methodologies within the community.

2.1.3 Feature Representation

An image is the representation of the external form of an object, generally as a composition of different regions. It is commonly agreed that feature extraction should not be merely assembling regional descriptors together. Concatenation of descriptors is the least favourable method, as it leads to high-dimensional features, and exacerbates the curse of dimensionality.

The most widely adopted alternative method in the literature is the “Bag of Visual Words” representation. To represent an image using this method, one should think of images as documents. Similar to documents, words in an image also need to be described. To describe these words, the steps of feature detection, feature representation and code generation have to be carried out. Feature detection is about extracting a number of local regions or patches. These are considered as basic elements in an image or also known as words in the described context. Feature representation deals with the problem of describing local patches as numeric values. Code generation is the

final step in the “Bag of Visual Words” model. This step involves converting vector represented patches to codewords. Each patch in an image is linked to a particular codeword through a clustering process. Eventually, the image will be represented by a histogram of codewords.

Aside from the “Bag of Visual Words” representation [13, 25], there exist alternative techniques like the covariance matrix representation [26, 27], fisher vector representation [16], and graph representation [28]. Bag of features discards all information about the geometry of underlying objects in an image. There is always a trade-off between perspective invariance and discriminative power. Therefore, preserving at least an approximation of image layout seems sensible for many classification problems. In reality, feature representations that benefit from some degree of spatial information, such as HOG [9], spatial pyramids [29] and their variants usually perform better in classification problems than pure “Bag of Visual Words” techniques.

2.1.4 Feature Combination

There exists an extensive body of literature on features in computer vision, and there are still new techniques emerging. However, it is evident that there are many problems that cannot be solved by a single ideal feature, and a combination of dissimilar types may be necessary. Different feature types capture different characteristics of an image. It is usually essential to combine various types of features to achieve a comprehensive understanding of an image.

The classic step after feature extraction in an image understanding problem is classification. Thence, the literature has various examples of feature combination techniques both at input and output level of classification step. For instance, the authors in [30] illustrate that outstanding semantic segmentation results can be achieved by simply concatenating local regional features with global bag of visual words descriptors as input to classifiers, whilst research in [31] proves that aggregating classifiers’ probabilistic outputs is a robust method of combination and performance gain.

Apart from simple aggregation techniques, kernel methods that represent relations between samples of different feature channels have gained popularity in the literature. This descriptor combination is performed at kernel level, which is regarded as a middle level fusion stage. More formally, let $\{x_i\}_{i=1}^N$ be N instances, and let $\{f_m\}_{m=1}^F$ be a set of F extracted features. Let us assume that the kernel function K_m is performed on the m^{th} feature channel. The similarity between two instances based on their m^{th} feature f_m is therefore defined as:

$$SIM(i, j) = K_m(f_m(x_i), f_m(x_j)) \quad (2.1)$$

Feature combination based on kernels is thus about combining different K_m into a single kernel K^* . For example, authors in [32] introduce a method beyond simple arithmetic to combine the kernels that produces superior results despite its simpleness. Other methods include the straightforward linear combination of kernels, which is amongst popular solutions in the literature:

$$K^*(x_i, x_j) = \sum_{m=1}^F \beta_m K_m(f_m(x_i), f_m(x_j)) \quad (2.2)$$

where β_m are the linear combination coefficients that can be learned by classification algorithms, if desired.

2.1.5 Feature Selection

In machine learning and statistics, feature selection is the process of selecting a subset of relevant features for use in model construction. This process is also known as variable selection, attribute selection or variable subset selection in the literature. The principal assumption in utilisation of a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those that provide no more information than the currently selected features. Irrelevant features provide no useful information in any context.

The simplest feature selection algorithm is to exhaustively test each possible subset of features in order to find the set that minimises the error rate. This is indeed an extensive search of the space, and therefore it is computationally intractable for all but the smallest of feature sets. Furthermore, the choice of evaluation metric has a massive influence on this simple algorithm. More interestingly, there are attempts in the relevant literature that exploit higher order statistics to derive a discriminative subset of features [33, 34]. These techniques enable low dimensional representation of inputs. They also allow selected features to be well-separated.

In this thesis, we will present a novel feature selection method that employs the maximum mutual information criterion to develop a supervised feature selection method for object representation in classification applications. Our proposed methodology based on mutual information exploits Fano's inequality [35] in a similar manner as authors in [36, 37]. In [38, 39] mutual information is also employed in deriving supervised but part-based

representations of objects. However, their methods are focused on extracting informative features from objects rather than alleviating the curse of dimensionality by finding discriminative subspaces.

2.2 Classification in Computer Vision

Classification is a machine learning concept that is common amongst many semantic image understanding problems. Classification is essentially a pattern recognition problem that is concerned with assigning a label to an input value. A supervised classification algorithm learns to assign labels to objects by observing examples, also known as the training data. It is common to come across methods in the literature that transform the recognition task to a classification problem, despite the fact that these two are not entirely analogous. In an ideal recognition task, the ability of recognising unknown objects instantly is preserved but the classification paradigm undergoes a burdensome step of one-to-one comparison of all learned classes that may be considered as its obvious defect, in addition to its inability to recognise unseen classes.

In the rest of this section, we will review a number of classification methods that are widely used in the relevant literature. These classifiers are also essential in building practical human in the loop algorithms.

2.2.1 Bayesian Classifiers

The established Naive Bayes classifier falls into this category of classification algorithms [40]. A Bayes classifier is a simple probabilistic method based on the theory of Thomas Bayes¹ with strong and naive assumptions of independence. Another term used in literature for describing this probabilistic model is the “Independent Feature Model”. There are numerous work in the literature [41–44] about applications of Naive Bayes classifiers for solving computer vision tasks.

Formally, let $p(c|F_1, \dots, F_n)$ be the posterior probability of a sample belonging to a certain class given a set of descriptors. Using Bayes’ theorem, the posterior can be written as:

$$p(c|F_1, \dots, F_n) = \frac{p(F_1, \dots, F_n|c)p(c)}{p(F_1, \dots, F_n)} \quad (2.3)$$

¹Thomas Bayes was an English statistician, philosopher and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes’ theorem.

In practical implementations, there is only interest in the numerator of this fraction, since the denominator does not depend on c , and the values of the features F_i are given. Hence, the denominator is effectively constant. The numerator is equivalent to the joint probability model: $p(F_1, \dots, F_n, c)$. Applying the chain rule for repeated applications of conditional probability, and assuming that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$ given the category c , the joint model can be expressed as:

$$\begin{aligned}
 p(c|F_1, \dots, F_n) &\propto p(F_1, \dots, F_n, c) \\
 &\propto p(F_1|c)p(F_2|c)p(F_3|c) \dots p(F_n|c)p(c) \\
 &\propto \prod_{i=1}^n p(F_i|c)p(c)
 \end{aligned} \tag{2.4}$$

The naive Bayes classifier combines the naive Bayes probability model with a decision rule. One conventional method is to select the hypothesis that is most probable. This is the familiar maximum a posteriori or MAP decision rule. The corresponding Bayes classifier is therefore defined as:

$$y(f_1, \dots, f_n) = \arg \max_c \prod_{i=1}^n p(F_i = f_i|C = c)p(C = c) \tag{2.5}$$

The assumption on distributions of features is called the event model of the Naive Bayes classifier. For discrete features, multinomial and Bernoulli distributions are amongst popular choices. To deal with continuous data, a typical assumption is that the values are distributed according to a Gaussian distribution.

Kernel Density Estimator is a non-parametric way of estimating the probability density function of a random variable. KDE is essentially a data smoothing problem where inferences about the population are made, based on a finite data sample [45]. When combined with a Bayesian classifier, it can be used in a supervised learning method.

2.2.2 Support Vector Machine

A Support Vector Machine (SVM) in essence is a mathematical algorithm that benefits from four basic concepts: a separating hyperplane, a maximum margin hyperplane, a soft margin that allows correct classification of data that is not separable and a kernel that is basically a mathematical trick in form of a function to project data from a lower dimensional to a higher dimensional space.

In linear binary settings, linear SVMs [46] learn a hyperplane that separates the training data based on their corresponding labels. Given a training dataset D , a set of n points of the form:

$$D = \{(x_i, y_i) | x_i \in \mathbb{R}, y_i \in \{-1, 1\}\}_{i=1}^n \quad (2.6)$$

where the y_i is either 1 or -1 , indicating the class to which the point x_i belongs. The objective is to find the maximum-margin hyperplane that divides data points with different class labels.

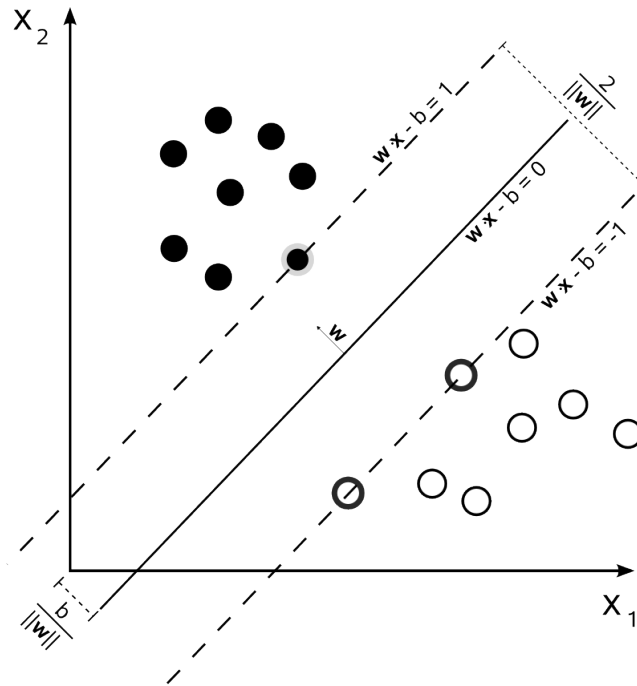


FIGURE 2.1: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

Any hyperplane can be written as the set of points x satisfying: $w \cdot x - b = 0$, where \cdot denotes the dot product, and w is the normal vector to the hyperplane. The parameter $\frac{b}{\|w\|}$ determines the offset of the hyperplane from the origin along the normal vector w . If the training data are linearly separable, two hyperplanes can be selected in such a way that no points lie in between them. The region bounded by these hyperplanes is called the “margin”. The distance between these two hyperplanes is $\frac{2}{\|w\|}$, thus it is required to minimise $\|w\|$ to maximise their distance. Figure 2.1 illustrates the resulting maximum-margin hyperplane and the support vectors. To prevent data points from falling into the margin, the following constraints are imposed for each i :

$$w \cdot x_i - b \geq 1 \quad \text{for } x_i \text{ of the first class} \quad (2.7)$$

$$w \cdot x_i - b \leq -1 \quad \text{for } x_i \text{ of the second class}$$

This can be rewritten as:

$$y_i(w \cdot x_i - b) \geq 1 \quad \text{for all } 1 \leq i \leq n. \quad (2.8)$$

SVMs can also handle non-linear data by adopting the kernel trick. A kernel $K(x_1, x_2)$ is defined as the inner product of functions $\phi(x_1)$ and $\phi(x_2)$:

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \quad (2.9)$$

where $\phi(\cdot)$ is a mapping function that projects the original data x into a higher (infinite) dimensional space $\phi(x)$.

For multiclass problems, x together with its label information y is projected into a joint high dimensional space $\phi(x, y)$. A standard multiclass kernel SVM can be defined as:

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & (w \cdot \phi(x_i, y_i) + b_{y_i}) - (w \cdot \phi(x_i, y) + b_y) \geq 1 - \xi_i, \quad \forall i, y \neq y_i \end{aligned} \quad (2.10)$$

where b is a vector composed of $\{b_y, y \in Y\}$. The non-negative slack variables ξ_i measure the degree of misclassification of the data x_i , and the optimisation becomes a trade off between a large margin and a small error penalty [47]. The predicted class y for a test sample x is then:

$$\begin{aligned} y &= \arg \max_{y \in Y} w \cdot \phi(x, y) + b_y \\ y &= \arg \max_{y \in Y} \sum_{i=1}^N \alpha_{iy} K(x_i, x) + b_y \end{aligned} \quad (2.11)$$

where α_{iy} is a Lagrange multiplier.

It is important to mention that Support Vector Machines generate uncalibrated values that are not probabilities. A sigmoid function is usually used to map the output of SVMs into probabilities. Support Vector Machines plus the sigmoid function will preserve the sparseness of SVM, while still yielding probabilities of useful quality [48].

2.2.3 Multiple Kernel Learning

Support Vector Machines deploy a single kernel matrix. There are many occasions where one kernel is either inadequate or the range of choice for kernels' types is very vast. Feature combination is an instance of such scenarios. Multiple Kernel Learning (MKL) is therefore defined as a methodology that aims to learn an optimal combination of different kernels. The following are examples of MKL implementations within the community.

The pioneering work of Multiple Kernel Learning dates back to Lanckriet et al. attempt in [49]. Nevertheless, the survey in [50] lists a number of recent MKL learning algorithms, and concludes minimum contrasts between them in terms of their accuracies based on a substantial empirical comparison. The formulation of MKL is still under scrutinisation by researchers, despite its success in many applications [51–55].

Typical MKL is essentially a linear combination of different kernels. This proves to be an unnecessarily strong constraint, as argued in [56]. The authors instead propose to learn augmented coefficients for every sample in each feature channel. This is achieved by augmenting the kernel matrices. Since the augmented kernel is a block diagonal matrix, the coefficients learned are equivalent to learning different kernels separately, and subsequently inserting a suitable bias term for all the kernel classifiers.

OBSCURE [52] is an alternative state-of-the-art multiclass multiple kernel learning algorithm that obtains excellent performance in a considerably lower training time. The conventional formulation of MKL algorithms is extended to accommodate a parameter that enables sparsity of a solution to be selected. The new proposed setting facilitates a fast convergence rate at lower iterations, as the number of kernels increases.

The MKL methodology proposed in [55] employs a state-of-the-art classifier to search for an object in all possible windows of an image. It uses the multiple kernel learning algorithm of Varma and Ray [51] to learn an optimal combination of exponential Chi-Square kernels. Each of these kernels captures a different feature channel. Such a powerful classifier to test all image sub windows is not efficient. Thus, the model introduces a three stage classifier that integrates linear, quasi-linear and non-linear kernel SVMs. The non-linearity of kernels increases their discrimination power at the cost of computational complexity.

The computational infeasibility is also aggravated by the fact that the number of regions to be searched in each image is large and feature histograms that describe them are high dimensional. To overcome this issue, their solution adopts a cascade approach, a multi-stage classifier, where each stage utilises a more powerful and more expensive

classifier. The first stage classifier employs fixed aspect windows [9, 57, 58], the second classifier considers multiple aspect ratios learnt from data and the third is a jumping window [59] classifier. The output of these classifiers is a set of candidate regions that are passed onto more powerful classifiers at the later stages.

Kernel-based methods, including SVM and MKL algorithms, are usually more robust than their linear counterparts. However, they are not scalable to large scale settings. On the contrary, resorting to linear algorithms leads to failure in dealing with non-linear data. This trade-off has fuelled the pursuit of alternative solutions in the community.

2.2.4 Artificial Neural Networks

The term “Neural Network” has its origins in attempts to seek mathematical representations of information processing in biological systems. Since then it has been used broadly to cover a wide range of different models [60–62]. Many of these models have been the subject of exaggerated claims regarding their biological plausibility. Nonetheless, biological realism would impose entirely unnecessary constraints from the perspective of practical applications of pattern recognition. Our focus in this thesis is therefore on neural networks as efficient models for statistical pattern recognition. In particular, we shall restrict our attention to the multilayer perceptron class of neural networks that possess very beneficial practical values.

The perceptron algorithm, also termed the single-layer perceptron, is the simplest feedforward network and a linear classifier in the context of neural networks. More formally, the perceptron is an algorithm for learning a binary classifier, a function that maps its input x to an output value $f(x)$:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.12)$$

where w is a vector of real-valued weights, $w \cdot x$ is the dot product $\sum_i w_i x_i$, and b is the bias, a term that shifts the decision boundary away from the origin and does not depend on any input value. Spatially, the bias alters the position not the orientation of the decision boundary. Formally, the quantities a_j known as activations of neurons are:

$$a_j = \sum_i w_i x_i + b \quad (2.13)$$

where each of these quantities are transformed using an activation function $h(a_j)$. A perceptron is an artificial neuron which deploys the Heaviside step function as the activation function.

The perceptron learning algorithm does not terminate if the learning set is not linearly separable. The most famous example of the perceptron's inability to solve problems with linearly nonseparable vectors is the Boolean exclusive-or problem. However, a modification of the standard linear perceptron can distinguish data that are not linearly separable.

A multilayer perceptron is a feedforward artificial neural network model that consists of multiple layers of nodes in a directed graph, where each layer is fully connected to the next one. Apart from input nodes, every node in this model is a processing element with a nonlinear activation function. The nonlinear activation functions enable the ability to distinguish data that are not linearly separable. They are generally chosen to be sigmoidal functions such as the logistic sigmoid or the “tanh” function. The overall network function for sigmoidal output unit activation functions takes the form:

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_j w_{kj} h \left(\sum_i w_{ji} x_i + b_j \right) + b_k \right) \quad (2.14)$$

where the set of all weight parameters have been grouped together into a vector \mathbf{w} . Thus the neural network model is simply a nonlinear function from a set of input variables $\{x_i\}$ to a set of output variables $\{y_k\}$ controlled by a vector \mathbf{w} of adjustable parameters. The multilayer perceptron encloses an input layer, an output layer and one or more hidden layers of nonlinearly activating nodes. Each node in one layer connects with a certain weight w_{ij} to every node in the following layer.

The multilayer perceptron utilises a supervised learning technique called backpropagation for training the network [62, 63]. Backpropagation, an abbreviation for “backward propagation of errors”, is typically employed in conjunction with an optimisation method such as gradient descent [64]. This method calculates the gradient of a loss function with respects to all the weights in the network. The gradient is fed to the optimisation method, which in turn exploits it to update the weights in an attempt to minimise the loss function.

Multilayer perceptron models were a popular machine learning solution in the 1980s, finding applications in diverse fields such as speech and image recognition. They faced strong competition from the much simpler support vector machines in the 1990s until

recently, where there has been renewed interest in backpropagation networks due to the successes of deep learning in various applications [22–24, 65, 66].

2.2.5 Ensemble Classification

Ensemble methods in the area of machine learning use multiple learning algorithms to obtain an improved predictive performance. This improvement is in comparison to what could be achieved from any of the constituent learning algorithms. The concept of combining classifiers is analogously proposed as a new direction for the improvement of individual classifiers' performance. These classifiers could be based on a variety of classification methodologies, and may achieve different rate of correctly classified individuals. Thus, the goal of classification result integration algorithms is to generate more certain, precise and accurate outcomes.

Numerous methods have been suggested for the ensemble of classifiers creation: using different subset of training data with a single learning method, using different training parameters with a single training method, and using different learning methods are all amongst these methodologies. Bootstrap aggregating (Bagging) [67] and boosting [68] are both commonly used techniques of the first category. In the rest of this section, we will review ensemble of decision trees and naive Bayesian classifiers as instances of methods that exploit the bagging technique.

2.2.5.1 Random Forest

One of the influential works on decision trees is the Classification and Regression Trees (CART) book of Breiman et al. [69], where the authors define the basics of decision trees and their application in classification and regression. Nonetheless, training optimal decision trees from data has been a long standing problem, for which one of the most prevalent algorithms is “C4.5” of Quinlan [70]. A number of years has passed since the introduction of decision trees. Their recent revival is due to the discovery that ensembles of slightly different trees tend to produce superior accuracies on previously unseen data. This is a phenomenon known as generalisation [71, 72]. Ensembles of trees will be discussed in this section but let us first focus on individual trees.

A tree is a collection of nodes and edges organised in a hierarchical structure. Nodes are divided into two groups: internal (split) and terminal (leaf) nodes. In contrast to graphs a tree does not contain loops. A decision tree is a tree used for making decisions. A decision tree can be interpreted as a technique for splitting complex problems into a hierarchy of simpler ones. From a high-level point of view, the functioning of decision

trees can be separated into offline and online phases that corresponds to training and testing stages respectively.

Random forest is an ensemble of random decision trees [72]. In typical classification settings, each tree in the forest is trained independently on a random subset of training data. For classification, the outputs from individual trees are combined to form the final prediction.

Formally, consider a set of training samples $\{x_i\}_{i=1}^N$, their corresponding class labels, and a set of extracted features as $\{F_i\}_{i=1}^N$. Split functions within the nodes of a tree divide the samples into two subsets, named left and right child respectively. Amongst practical splits are the familiar linear classifiers:

$$\begin{cases} w^T F + b \geq 0 & \text{go to left child} \\ \text{otherwise} & \text{go to right child} \end{cases} \quad (2.15)$$

Multiple splits are generated by selecting different feature dimensions or thresholds. The widely used information gain criteria [73–75] is typically used to find the best split at each node in the training stage:

$$\text{Score}(\text{split}) = \Delta E = -\frac{|G_l|}{|G_n|}E(G_l) - \frac{|G_r|}{|G_n|}E(G_r) \quad (2.16)$$

where G_n is the set of training samples in node n . $E(G)$ is the Shannon entropy of the class distribution in the set of samples G . G_l and G_r represent the training images contained in node n 's left and right child node respectively. The Shannon entropy is defined mathematically as:

$$E(G) = -\sum_{c \in C} p(c) \log(p(c)) \quad (2.17)$$

where C is the set of class labels for samples G . This split action is performed recursively on training samples until the stopping criterion is satisfied. The criterion could be the maximum depth of the tree. Alternatively a minimum information gain can be imposed. Tree growing may also be stopped when a node contains less than a predefined number of training points. Avoiding full grown trees has been demonstrated to have positive effects in terms of generalisation [71, 72].

During training, information that is useful for prediction in testing will be learned for all leaf nodes. In a classification scenario, each leaf may store the empirical distribution

over the classes associated to the subset of training data that has reached that leaf. The probabilistic leaf predictor model for the t^{th} tree is therefore: $p_t(c|x)$. To classify a test sample as some c class, it is sent through all trained trees. It starts from the root node and traverses down to right or left nodes based on the split function, and eventually falls into one of the leaf nodes. Leaf nodes store a normalised probability distribution of the occurrence of all possible classes in the dataset. Tree testing can often be performed in parallel, thus achieving high computational efficiency on modern parallel CPU or GPU hardware.

The outputs from different trees are combined together as the final result of the random forest. Combining all tree predictions into a single forest prediction may be carried out by a simple averaging operation. For instance, in classification final output is obtained by:

$$p(c|x) = \frac{1}{T} \sum_{t=1}^T p_t(c|x) \quad (2.18)$$

where T is the number of decision trees in the forest. Although trees are not statistically independent, it is possible to alternatively multiply the trees' outputs together to form a final result:

$$p(c|x) = \frac{1}{Z} \prod_{t=1}^T p_t(c|x) \quad (2.19)$$

where the partition function Z ensures probabilistic normalisation.

Previous research on random forest has commonly focused on its discriminative power. For instance, it has been successfully deployed in applications such as: image classification [75], object detection [76], and human pose estimation [77]. Recent years have seen an explosion of forest-based techniques in the machine learning, vision and medical imaging literature [74, 75, 78–82]. A recent success story of applying decision forests in a practical computer vision setting is perhaps the Microsoft Kinect for Xbox 360 console [77, 83].

2.2.5.2 Random Naive Bayesian

Randomised learning methods rely on two major concepts to train an ensemble of similar type classifiers: i) employing random input selection, and ii) random feature selection. The main advantages of these methods are the improved stability and decreased variance

of the resulting classifier. Using random input selection techniques, such as bagging, several classifiers are trained on different subsets of the training space and additively combined to form an ensemble. Bagging improves the final classifier in terms of stability and classification accuracy. It also helps to avoid overfitting.

The low computational and memory costs of random naive Bayes classifiers makes them suitable for applications where computational power and memory are limited or if very large datasets have to be processed. Interestingly, the work from Prinzie and Van den Poel in [84] illustrated the fact that the tree structure of random forests can be replaced by simpler learning methods such as naive Bayes without a significant loss in the performance. For instance, authors of [85] developed an efficient online learner by adapting the random naive Bayes classifier to the online domain. They propose to use online histograms as weak learners, which perform superiorly compared to simple decision stumps. Their approach is applicable to incremental learning on machine learning datasets, and it is empirically evaluated on the task of tracking by detection.

In short, the random naive Bayesian classifiers ensemble is created by randomly selecting F features out of the feature pool D . The class conditional probability distribution $p(x_f|y)$ of every feature is subsequently modelled for each class y . The probability of a sample observation x belonging to the class y can then be described by combining B randomly trained naive Bayes classifiers, each using $F \leq |D|$ features.

2.3 Understanding Visual Content at Pixel Level

Thus far, we have reviewed the most significant building blocks of many computer vision algorithms: features, and classifiers. We now review their utilisation for the purpose of semantic understanding. The literature in visual content understanding at pixel-level is conventionally divided into two broad categories: object detection, and image segmentation. The following are amongst prime examples of work in these categories.

2.3.1 Object Detection

Object detection is usually performed either for rigid objects with specific shapes, or those of amorphous spatial extent. Examples of the former are pedestrians, and cars whilst trees, road, and sky are amongst instances of the latter. Most current methodologies in the literature are devised to deal with rigid objects, since it is still very challenging to detect non-rigid objects.

The most straightforward approach in object detection is the sliding window method. A predefined sub-window slides over the test image to cover all possible locations and scales. The features in the sub-window are extracted, and fed to a classifier that has been trained to determine whether the sub-window contains the specific object or not. This method is amongst the dominating approaches despite its simplicity.

For instance, the authors of [86] address the problem of object detection in a unique way. Their underlying model utilises a graph whose nodes represent a dense set of regions. The edges of the graph illustrate the grid structure of the underlying image. They act as springs to ensure that the geometry of nearby regions remains consistent during the matching process. The algorithm constructs a kernel suitable for SVM based classification using only one type of feature. The matching process formulates an energy optimisation problem defined over graphs with a coarse grid of the underlying image. Their framework for image classification can be readily extended to object detection using sliding windows.

Recent modifications to the sliding windows method attempt to increase the detection efficiency using branch and bound [87], or to combine a holistic window with inner parts of the sub-window that represent elements of the object in test [88]. However, a prominent example of a state-of-the-art detection system is perhaps the deformable part-based model of [89]. It builds on carefully designed representations and kinematically inspired part decompositions of objects, expressed as a graphical model. Using discriminative learning of graphical models allows them to build high-precision part-based models for variety of object classes.

In real world scenes, objects are often occluded by other objects. Hence, part-based methods are inherently more robust than holistic approaches in dealing with occlusion in real world applications. Current influential methods include the constellation model of [90, 91], and the Hough voting based techniques in [92]. The constellation model treats the object as a constellation of local parts, and infers their optimal combination in a Bayesian framework. Albeit the strategy of treating an object as a constellation of parts is more robust to occlusion, its defect is its lack of discriminative power.

It is commonly understood that human object recognition systems rely heavily on context. Hence, many existing object detectors in the literature exploit context to improve their performances. For instance, work in [93] benefits from image categorisation as a prior to guide object detection, while [94] deploys an entropy criterion to select unknown objects. Concatenated features of neighbouring known classes and unknown objects are then used to detect the remaining unknowns.

In summary, object detection is used to detect only a certain class in an image. However, it remains an unresolved problem despite the fact that there are numerous attempts in the literature aiming to tackle the issue. The state-of-the-art techniques, such as [89], are still below real world requirements in some applications.

2.3.2 Image Segmentation

In computer vision, image segmentation is the process of partitioning an image into multiple segments or a set of pixels. The primary reason for this task is to simplify an image into something that is more straightforward to analyse. Image segmentation is normally deployed to locate objects or find boundaries within an image.

The objective of [95] is the unsupervised segmentation of image sets into background and foreground. The resulting segmentation will improve the classification performance. Their algorithm has better performance compared to many of its predecessors due to the fact that the actual segmentation task is carried out at different levels: pixels and colour distributions for individual images and super pixels with learnable features at image set level. These levels together with powerful inference algorithms, such as SVMs, result in the high performance.

The practical importance of segmentation is observed in [96], where the authors consider an automated processing pipeline for tissue micro array analysis, also known as TMA, of renal cell carcinoma. The tasks to achieve such analysis map to several challenging machine learning challenges such as nuclei segmentation and classification. The segmentation of cell nuclei is performed using Graph Cut [97–99]. Several shape and histogram features are extracted from the resulting segmentations. To achieve reliable classification results, an SVM with different kernels and distances is used. The results illustrate that all extracted features from segmentations are essential to an optimal performance. The carefully selected kernels perform considerably better than chance and are analogous to human domain experts.

Table 2.1 illustrates four different commonly used segmentation algorithms and their employed methodologies. Interactive Graph Cuts and Binary Partition Trees have proved to be the most effective algorithms in terms of average accuracy over time [100].

2.4 Understanding Visual Content at Image Level

The idea of understanding the semantic meaning of each pixel may seem very appealing. However, this has proved to be very difficult to achieve at current levels of technology.

TABLE 2.1: Segmentation Algorithms

Methodology	Algorithm
Region Growing	Seeded Region Growing [101]
Graph and MRF Models	Interactive Graph Cuts [98]
Classifiers	Simple Interactive Object Extraction [102]
Hierarchical, Split and Merge	Interactive Seg. using Binary Partition Trees [103, 104]

It is therefore suggested to take a step back, and contemplate on a more abstract level of semantic image understanding. There even exist scenarios where we only need image-level semantic meanings. The Content Based Image Retrieval (CBIR) systems are instances of such situations. A precise image-level semantic understanding can even serve as context information, which in turn may enhance the performance of semantic understanding at pixel-level [93].

The rest of this section lists three major tasks that are concerned with visual content understanding at image-level: content classification, content based image retrieval, and automatic annotation. These three tasks are indeed closely related. An image annotation algorithm could automatically assign an image with a number of keywords. An image retrieval system then retrieves images directly based on these suggested keywords. Furthermore, the problem of image annotation itself could be decomposed into a set of image classification tasks, where each task aims to predict the existence of a distinct tag.

2.4.1 Visual Content Classification

The literature in the field of visual content classification is extremely rich, and it is infeasible to cover all aspects of such a vast domain. Alternatively, we aim to select three key subproblems related to the field, and explore their details:

1) The vision task of **multiclass classification** becomes challenging when the number of classes is very large. Testing against every single class is computationally infeasible in such cases. This complication can be solved by learning or imposing a structure over the set of classes.

The solution presented in [105] introduces a method for fast multiclass classification by learning label embedding trees and optimising overall tree loss. The proposed solution is faster than One-vs-Rest methods while yielding comparable accuracy to state-of-the-art frameworks. Their approach relies on two main ideas. Firstly, each node in the label tree predicts the subset of labels to be considered by its children. This will decrease the number of label classes at a logarithmic rate until a prediction is made. Secondly,

both image features and labels are jointly embedded into a low dimensional space using a linear transformation, which still optimises the overall tree loss.

As a practical application of multiclass classification, a novel approach to natural scene categorisation is introduced in [106], where a collection of local regions in a scene is represented as part of a theme. Such themes are conventionally learnt from hand annotations of experts. In contrast to the norm, this model learns such themes without supervision. It learns characteristic intermediate themes of scenes without any human intervention. The model is capable of categorising images into hierarchies by a Bayesian framework, yielding a performance similar to what humans can do under normal circumstances. The algorithm in their solution is based on a Latent Dirichlet Allocation model introduced by Blei et al. in [107].

2) In spite of complications caused by a large number of classes in multiclass classification scenarios, computer vision research has seen great success in basic level categorisation. This is in contrast to **fine-grained categorisation**, which has received little attention. These are classes of objects which are not usually recognisable by ordinary human beings. Simple examples of such classes may range from animal species and aircraft models to botany [108, 109]. Unlike basic level categorisation, fine-grained categorisation may be problematic even for humans. Thus, an automated system to accomplish this task could prove valuable in many applications.

The goal of fine-grained categorisation is to some extent achieved in [110] by discriminative feature mining and randomisation. The latter is essential to be able to handle the massive feature space and prevent the problem of overfitting. Their model proposes a random forest [72] with discriminative decision trees to determine image patches that are highly discriminative for the purpose of categorisation. Each decision tree in the forest considers a small number of patches. This will ensure little correlation between trees and consequently better performance on their fine-grained image classification problem.

3) An ideal algorithm of visual content classification needs to possess the ability to distinguish between known objects and those of unseen categories. Hence, the goal of **object category discovery** is to automatically identify groups of image regions that belong to some unknown category. This task is usually performed in a purely unsupervised setting. Therefore, the performance of such categorisation depends on accurate assessments of similarity between unlabelled regions in images.

Authors in [111] introduce a new framework to cluster unlabelled data more accurately by learning from a set of labelled categories and optimising similarity. Their model demonstrates that including both labelled and unlabelled training data, when optimising

similarity metric, leads to general improvement in terms of quality of the system. Their implementation includes a multiple kernel, and an optimised similarity space. Image segmentations are classified into familiar and unfamiliar by a k-nearest-neighbour algorithm. Unfamiliar segmentations are clustered in the optimised space in order to enable discovery of new categories.

2.4.2 Content Based Image Retrieval

Content based image retrieval (CBIR), also known as query by image content (QBIC) and content based visual information retrieval (CBVIR), is the application of computer vision methodologies to the image retrieval problem. Fortunately, there exist useful review surveys [2, 112] that list references to a huge number of systems and their core technologies in the field. In simple words, CBIR systems like [113, 114] propose novel ways to search for digital images in very large databases.

In CBIR systems, many feature extraction strategies have been proposed for retrieving images that are similar in terms of colour, texture, shape, etc. However, it is important to note that features that are effective for classification problems may not be suitable for retrieval and display of visually similar images, in particular for medical CBIR systems. Applications of CBIR and classification in medical imaging have already been presented in the literature but they are mostly targeted at radiological images [115].

In an interesting approach [116–118] present CBIR systems as a diagnostic tool for skin lesion photographic images. They illustrate that using composite features improves the overall performance, in comparison to the utilisation of a larger number of standard features. A genetic algorithm is used to combine simple features using a series of operations in order to derive the synthesised descriptors.

2.4.3 Automatic Image Annotation

As the name is self-explanatory, proposed solutions of image annotation problem attempt to find practical ways to automatically annotate image content with meaningful keywords. For instance, the “Relevance Model” estimates the joint probability of keywords and the image [119].

A simple framework for automatic image annotation using global features and robust non-parametric density estimation is presented in [120]. Their model employs the Bayes’ theorem to invert the conditional dependencies of choosing a suitable word in regards to a given image. The method of inference in the proposed model is a non-parametric density estimator formulated by the kernel smoothing technique of Parzen [45]. It is

known that smoothing improves efficiency for finite samples. Global colour features are used for modelling keyword densities and the popular Earth Mover's Distance metric is also effectively incorporated within this framework.

Most previously proposed annotation methods assign keywords separately but the correlation between keywords to improve image annotation performance has recently received considerable attention. Nevertheless, estimating the joint probabilities of sets of keywords and unlabelled images has proved to be computationally unmanageable. In order to overcome the issue of computation, [121] proposes a heuristic greedy iterative algorithm to calculate the probability of a suitable keyword as a semantic caption of an image. In their approach, the correlations between keywords are analysed by "Automatic Local Analysis" of text information retrieval. In addition, a new image generation probability estimation method is proposed by them based on region matching.

The co-occurrence of keywords in a limited training set is rare, which translates into a very sparse co-occurrence matrix. Nonetheless, a zero probability in the training data does not necessarily mean the correlation will never occur in the future. Therefore, various smoothing methods have been introduced in the relevant literature to rectify the problem. Non-negative Matrix Factorisation [122] and the Jelinek-Mercer algorithm [123] are two examples of such smoothing methods. They allocate a small non-zero probability to the keywords in the matrix.

A hybrid probabilistic model is introduced in [124] to solve the problem of content based image tagging. The proposed solution integrates low-level image features and high-level user provided tags to automatically annotate images. The approach exploits a tag-image association matrix. The number of images is usually very large and user provided tags are very diverse in this matrix. This means that the association matrix is very sparse and difficult to be used directly for estimating tag to tag co-occurrence probabilities. Thus, they introduce a collaborative filtering method based on Non-negative Matrix Factorisation to tackle the issue of data scarcity. An L_1 norm kernel method is employed to estimate the correlations between low-level image features and semantic concepts provided by human knowledge.

Supervised learning from multiple sources of annotation data has been a challenging problem to this date. Combining knowledge from different information sources is far from being a solved problem. The increasing availability of more annotators from different expertise domains, the difficulty of obtaining ground truth in particular cases such as cancer detection in medical images, in addition to the subjectivity of annotators portray the importance of studying supervised learning when there are multiple annotators with variable skills. The labels annotators provide may be unreliable, noisy or inconsistent depending on the instance of data they observe. The projected system in [125] develops

a probabilistic approach to the problem of annotators from different sources with various levels of expertise. Their model is suitable for dealing with missing annotators, estimating the ground-truth, and evaluation of annotators. The implementation is carried out in the context of statistical inference once the correct conditional distribution is observed. Their presented approach produces classification and annotator models that allow estimates of the true labels and annotator variable expertise.

The goal of finding a solution to the problem of image annotation, where class labels are not easily attainable, has been explored in [126]. The authors of this work propose to utilise tags of training images as the supervising information to guide the generation of random trees. This enables the retrieved nearest neighbour images to be not only visually alike but also semantically related. It is important to mention that unlike the conventional decision forests, which fuse the information contained at each leaf node individually, their method treats the random forest as a whole and introduces new concepts such as: “Semantic Nearest Neighbour (SNN)” and “Semantic Similarity Measure (SSM)” that approximately indicate “which” and “how many times” training images fall on the same leaf node with the query image. Succinctly, they annotate an image from the tags of its semantic nearest neighbour based on their proposed semantic similarity measure. Their new technique is intrinsically scalable and competitive to state-of-the-art methods. In the upcoming chapters, we will introduce a method of reducing strain imposed on users of human in the loop applications based on this adoptable annotation framework.

2.5 Understanding Visual Content by Human in the Loop

Many problems in the field of computer vision are immensely difficult or even impossible to be solved entirely by automatic solutions. Thus, in tackling such problems, it is not only helpful but also vital to explicitly incorporate high-level knowledge of humans and their intentions. The fundamental technical challenges based on this philosophy are therefore to capture and harness such abstract knowledge computationally. Human in the loop algorithms can be utilised to represent interactive, hybrid human-computer methods for the purpose of object classification, segmentation, and annotation in computer vision settings.

2.5.1 Visual Content Classification

In traditional passive approaches, the output of an object detector is combined with high-level knowledge in a post processing phase in order to boost classification

performance of a particular problem such as scene recognition. In contrast, a new framework in [127] demonstrates an active approach that benefits from high-level knowledge by implementing an interaction between a reasoning module and a sensory module. The reasoning module gathers high-level knowledge about a scene and its object relations. It also commands the sensory module to alter its attentional focus and determine the contents of the scene. On the other hand, the sensory module is responsible for detecting objects and extracting features from images. The novelty of such an active paradigm is that the sensory module, guided by the reasoning module, shifts its focus of attention to a small number of objects in the scene. This translates into faster and more accurate scene recognition. The attention mechanism is achieved by using a maximum information gain approach. This means that each detected object should maximise the augmented information for scene recognition.

Visual attributes is another interactive solution that provides a beneficial intermediate representation between low-level image features and high-level categories of classification problem. These attributes are gaining importance in the recognition literature. Attributes that are both nameable and discriminative appear to be in disagreement. The paper in [128] introduces a method to define a vocabulary of attributes that is simultaneously nameable to humans and discriminative to machines. Their model demonstrates a way to actively augment this vocabulary with new attributes that resolve confusion at class level. The framework also proposes a novel way to prioritise candidate attributes by their probability of being associated with a nameable property. The key technical obstacles solved in this model are determining attributes based on visual features separability and current class confusion, modelling the nameability of such attributes and finally selecting representative image examples that will prompt reliable human responses of attribute names.

Lampert et al. in [129] study the classification problem when training and testing sets are disjoint. This is when there are no training examples of the target classes available. They attempt to solve the problem of object classification by utilising human specified high-level description of target objects rather than training images. The description contains arbitrary semantic attributes such as shape, colour and even geographic information. The proposed method solves the problem by transferring information between classes. This transfer is by means of an intermediate representation that consists of high-level and semantic per class attributes. This method facilitates an efficient way to incorporate human knowledge into the system. In the proposed multiclass attribute based classifier, the posterior distribution of the training class labels at test time derives a distribution over the labels of unseen classes by using an intermediate class attribute relationship.

From a different perspective, the proposed model in [130] utilises a similar technique to the 20 questions game where the visual content of an image interactively poses the next possible question. The goal is to correctly classify the object with a minimum number of questions asked. Their methodologies account for imperfect user responses and unreliable computer vision algorithms. User inputs in the system raise the accuracy to levels that are practical for an application, whilst at the same time reducing the amount of time and human interaction required.

Their model exploits visual content of an image and the current history of question responses to intelligently ask the next question. Maximum information gain, which is widely used in decision trees [69], is used as the criterion to select the following question. Question responses are estimated as a multinomial distribution with parameters learnt from a training set of user responses collected from Mechanical Turk². The training set includes user responses along with their confidence in their answers. The confidence or certainty value is parametrised by three distinctive options: guessing, probably and definitely. The training set is also incorporated with a Dirichlet prior to improve robustness and performance when the training data is sparse. The model in this paper allows any off-the-shelf multiclass object recognition algorithm to be plugged into the visual 20 questions game. In their experiments, vision algorithms based on Andrea Vedaldi's publicly available source code [55] are utilised to evaluate the datasets.

A later work [131] from the same group approaches local part categorisation with the emphasis on users to locate different parts of an object. As in the previous example, this model is also designed for fine-grained visual categorisation. The machine intelligently asks the most appropriate question and the user responds by either answering the binary question or clicking on object parts. By employing computer vision algorithms and analysing user responses, the overall amount of human effort is minimised, while the accuracy results show an improvement over challenging datasets of uncropped images with noisy backgrounds. This achievement means that the proposed solution counts for errors and inaccuracies of vision algorithms and ambiguities in multiple forms of human feedback like their perception of part location, attributes, and corresponding class labels.

2.5.2 Image Segmentation

An interactive solution to the segmentation problem is presented in [132]. In their new region merging based method, users only have to approximately mark the location and region of objects and background by using strokes, also known as markers. A novel maximal similarity based region merging mechanism is introduced to direct the

²<https://www.mturk.com>

process. The proposed model automatically merges initial segmented regions by mean shift segmentation. It also defines object contours by labelling all non-marker regions as either object or background. The matching process is dynamic and adaptive to image content. Thus, it does not require a similarity threshold to be set in advance.

An important requirement in designing an interactive system is its usability. The authors in [133] present an interactive segmentation tool with such considerations in mind to ensure a high level of user experience. The proposed tool can rapidly and easily evolve optimal image segmentation parameters from scratch. It incorporates user local search and makes the fitness function more dynamic to enhance the underlying decision making process. Further improvements, such as a hybridising evolutionary algorithm, are made in the proposed framework. This algorithm contains domain specific knowledge in form of hint features. These features guide the mutation or utilisation of more texture kernels to work with a wider range of images.

2.5.3 Semi-automatic Image Annotation

The image annotation framework in [134] proposes a solution to the problem of interleaving interactive labelling. The annotation of new examples is semi-automated by an online learning model, where a recently labelled example is used to update the system's parameters. The framework is specifically applied to solve the problem of part-based detection and interactive labelling of deformable part models.

This approach takes advantage of both strongly and weakly supervised methods. The former delivers computational tractability, whilst reduction in human interaction time is granted by the latter. Furthermore, online algorithms are employed to optimise a structured SVM function for incrementally training a vision system that is capable of doing more mundane or obvious labelling tasks. Their system gradually develops into a more powerful solution, which eventually reduces the amount of human annotation time per image.

A novel multilabel learning framework, also known as Semi-Automatic Dynamic Auxiliary-Tag-Aided, is presented in [135] to interactively solve the problem of image annotation. Their framework boosts the classification rate of a target tag by the classification results from a subset of other annotations, known as auxiliary tags. These auxiliary tags, which are strongly correlated with the target tag, are determined in terms of normalised mutual information. For a given image, the set of target tags recommended by the auxiliary classifier can be refined by user feedback and therefore the proposed model will try to suggest more appropriate tags in the next iteration of the algorithm. In contrast to traditional static methods, this speeds up the image annotation procedure.

2.5.4 Medical Applications of Human in the Loop

Medical applications of human in the loop settings are relatively limited in the computer vision literature. For instance, existing approaches to exploiting Information and Communications Technology (ICT) in dermatology, such as teledermatology (TD) and computer aided diagnosis (CAD) have had limited success [136]. TD's total reliance on human experts viewing electronic images from a remote location to perform disease diagnosis is severely hindered by a shortage of human specialists. The core technology for CAD, computer vision, is still an evolving research subject and performances are not yet practically useful. As a result, almost all research in applying CAD to dermatology has been limited to melanoma conditions and using clinical or dermoscopic images [136].

An example of a traditional computer aided diagnosis system is an automated melanoma recognition framework introduced in [137]. Initially, a binary mask of lesion is obtained by a number of basic segmentation algorithms alongside a fusion strategy. A set of shape and radiometric features is calculated to determine the malignancy of a lesion. As a different approach, a physics-based model of tissue colouration [138] provides a cross-reference between image colours and the fundamental histological parameters of skin lesions. The model is built by computing the spectral composition of light remitted from the skin. The model is representative of all the normal human skin colours. Abnormal skin colours do not conform to this model and thus can be detected.

The authors in [139] utilise optical spectroscopy and a multi-spectral classification scheme using SVMs to assist dermatologists in their diagnosis of skin lesions. Another solution is a computer image analysis system presented in [140] that differentiates early melanoma from benign pigmented lesions. The analysis system extracts features related to the size, shape, boundary, and colour of each lesion. Feature extraction in [141] is limited to the quantification of degree of symmetry. The symmetry quantification step presents a six dimensional feature vector that can be exploited to classify pigmented skin lesions as benign or malignant. The solution demonstrates that the underlying scheme outperforms methods based on the principal component decomposition that is generally used for this category of applications.

A more practical framework is proposed in [142] that assesses a series of 588 flat pigmented skin lesions. The proposed analyser groups 48 parameters into 4 categories that are used to train an artificial neural network. A feature selection procedure confirms that as few as 13 of the variables are adequate to discriminate the two groups of "melanoma" and "other pigmented" skin lesions.

Surprisingly limited research exists in applying computer vision techniques to recognising other common skin conditions based on ordinary photographic images. Furthermore, wide availability of mobile computing and smart phone devices have spurt extensive activities to exploit these technological advancements for dermatology applications. Carefully studying 79 dermatology-themed smartphone apps surveyed in [143] has intriguingly come to two conclusions: ubiquitous mobile computing technologies offer new opportunities and possibilities for developing new applications in dermatology to help improve patient care; however, all existing systems follow the traditional TD paradigm and none have intelligent CAD capabilities. These conclusions led us to take an interest in applying human in the loop algorithms on a dermatology dataset of various skin conditions, and not focus merely on methods of diagnosing melanoma cases from dermatoscopic images.

It is sensible to mention that there are examples of human in the loop approaches [144, 145] in alternative domains of medical imaging besides dermatology applications. Authors in [144] for example present an approach to Content Based Image Retrieval (CBIR), which combines the expertise of a human, image characterisation from computer vision, and automation made possible by machine learning. Although they have an overall classification accuracy of approximately 93%, this result is not uniform across disease classes. For less populous condition classes, their accuracy can be far lower. Their solution can also benefit from a better utilisation of user feedback when retrieval results are judged unsatisfactory. Authors of [145] introduce a physician in the loop content based image retrieval system for HRCT image databases that suffers from similar shortcomings. High-resolution computed tomography (HRCT) is computed tomography (CT) with high resolution, which is used in the diagnosis of various health conditions.

Although the literature demonstrates a number of attempts at fabricating CBIR medical systems for dermatological purposes [118, 146], and quite a few attempts at assessing severity of specific conditions automatically [147], the lack of a reliable medical system for unskilled users, who may provide misleading information, is apparent.

2.5.5 Information Source Fusion in Human in the Loop Applications

Information fusion in our context is the process of integration of multiple sources of data and knowledge representing the same object into a consistent, accurate, and useful representation. It consists in the merging of information in order to deduce a decision less noisy than the one obtained with only one source of information. Information fusion is an established area and data fusion systems are now widely used in various settings

such as: sensor networks [148], robotics [149], image processing [150], and computer security [151].

Information fusion is common in classification [152]. The process of fusing information sources in classification settings can be accomplished either at input or output end of classifiers. At input level, information sources in form of feature vectors are jointly fed into a classification algorithm. This is usually considered to be a method of data aggregation. As opposed to input level fusion, predictions from individual classifiers can be combined to produce a final result at output end of classification models. For instance, the problem of combining classifiers, which use a single source of information, has been previously studied in [31]. The authors introduce a common theoretical framework for combining classifiers by a number of simple schemes such as the product, sum, min, max, and median rule. They also compare these combination schemes to a majority voting strategy that assigns the class label based on the number of votes it receives from available classifiers. The sensitivity of these schemes to estimation errors is also investigated in their work to establish the fact that the sum rule is the most resilient combination scheme amongst the rest. Ensemble methods [153, 154], such as bootstrap aggregating (Bagging) [67] and boosting [68], are also widely used techniques of fusion at output level. State-of-the-art ensembles of convolutional neural networks [155] have recently made significant impacts on the challenging ImageNet computer vision competition [156].

Unlike classification, information fusion techniques have not been fully exploited in the context of human in the loop vision applications. The sources of data in human in the loop settings commonly include visual features extracted from images and high-level information obtained from users in the loop. Therefore, human in the loop approaches need to solve a supervised learning problem induced by abundance of choice in selecting the correct prediction from available classifiers that are trained separately on multiple sources of visual and user provided information.

One of the pioneering examples in fusion of information sources at output level for a human in the loop application is described in [130]. They propose to use the Bayesian framework to combine the visual information and user answers to perceptual questions for a bird species recognition system. It is our understanding that their method does not fully exploit the option of intelligent fusion for the two sources of low-level visual and high-level human provided information, as each of the sources in their framework is estimated separately and put together subsequently with equal weights to form an answer. This kind of later fusion is the norm in the literature [157, 158], although the lack of estimating interactions between visual features and user answers is known to be an issue.

Attempts to rectify the aforementioned problem of equal weight fusion are presented in [159–161]. In a slightly different domain, authors of [159] introduce a multimedia retrieval system that jointly models visual and textual components of a sample. As any other similar system, their human provided information is annotations from a range of users with inconsistent quality in their work, and thus the available annotations are not complete. An algorithm that evaluates the effectiveness of visual and textual components separately, and performs intelligent fusion is still desired.

To the best of our knowledge, research so far has not considered the problem of information source fusion for human in the loop classification applications in detail. In the following chapters of this thesis, we will emphasise the importance of intelligent information source fusion in enhancing efficacy of classification tasks by reviewing empirical results from a number of human in the loop datasets.

2.5.6 Current Shortcomings in Human in the Loop Approaches

We firmly believe that deep review of current methodologies in the human in the loop literature highlights several issues that still require a degree of contemplation:

i) It is of utmost importance to devise creative ways to harness human abstract knowledge in a manner that is beneficial to the computer vision algorithms. ii) It is of great interest to devise algorithms that find the most discriminative features amongst available descriptors that are either based on visual information or crafted from human abstract knowledge. iii) The burden on human in the decision making loop of a computer vision algorithm has to be kept at minimum, whilst the advantage of utilising their knowledge remains intact. iv) The fusion of low-level image information and high-level human knowledge demands robust generative or discriminative frameworks. v) It is preferable to have a fusion framework that intelligently selects the most reliable source of information available either from images or users in the loop.

In the next few chapters of this thesis, we will be scrutinising each of the aforementioned issues, and will suggest possible solutions to rectify them for realisation of robust human in the loop computer vision algorithms.

2.6 Summary

Visual content understanding is a vast field of research with numerous techniques and methodologies proposed in the literature, and it is infeasible to list all related works

in one section. Therefore, we listed the most relevant vision methods to the theme of human in the loop in this chapter.

We will further review related work to our proposed techniques and methodologies in the following chapters, where it deems necessary.

Chapter 3

Discriminative Object Recognition with Humans in the Loop

Human in the loop classification frameworks are typically designed to enhance recognition performance by fusing low-level visual information of images with high-level knowledge of users. Amongst the most intrinsic problems of these frameworks are the absence of techniques for representing human abstract knowledge, and the abundance of fusing methodologies for merging heterogeneous sources of information.

In this chapter, we introduce a discriminative random forest framework that harnesses human knowledge in a compact representation form and performs a simple but effective task of information source fusion. Our proposed approach reviews several core problems in realisation of discriminative human in the loop image recognition frameworks and attempts to answer these fundamentally relevant questions:

1. How to efficiently utilise user-provided information?
2. How to employ this abstract knowledge in an interactive fashion?
3. How to fuse low-level visual features and high-level user information for an enhanced recognition performance, subject to compatibility with discriminative classifiers like random forests?

We will evaluate our introduced solution on a number of compatible datasets, and as an application of the human in the loop approach, we aim to exploit our proposed interactive

methodology to build an innovative tool that assists primary healthcare workers with recognising various skin conditions.

Worldwide, it is believed that there are between 1000 to 2000 possible skin conditions and around 20% of them are difficult to diagnose [162]. Skin diseases have a major adverse impact on quality of life and many are associated with significant psychosocial mobility. A recent comprehensive assessment of healthcare needs for skin conditions in the UK [162] suggests that 54% of the population experience a skin condition in a given twelve month period and around 23% to 33% of the population have skin problems, which can benefit from medical care at any one time. The UK healthcare system relies on primary care as gatekeepers. Despite skin disorders being one of the most common reasons that people refer to their general practitioners, typical GPs paradoxically get minimal training in dermatology. Clearly, there is an acute skill shortage to meet the healthcare needs of the nation [162]. Furthermore, as the ratio of consultant dermatologists to the general population has remained very low in many resource-poor countries [163], a system that could automatically recognise skin diseases would be ideal to meet this apparently worldwide need.

A clinically approved tool would ultimately offer sufferers correct treatment and care in a timely manner; thus reducing their suffering, whilst enhancing the quality of their lives. Figure 3.1 illustrates our proposed human in the loop tool and its usability for recognition of skin conditions.

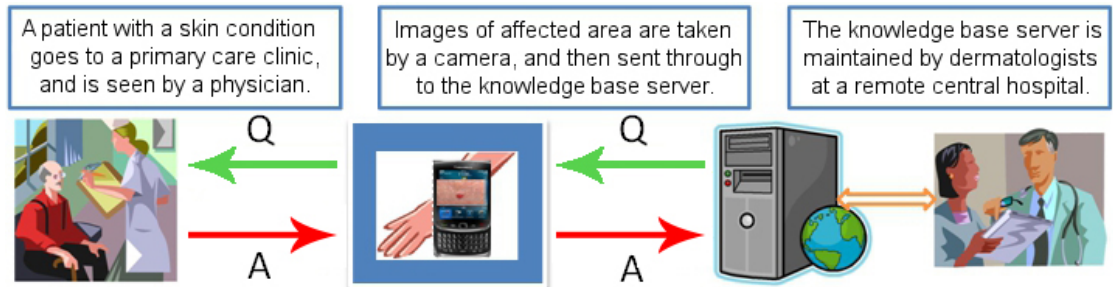


FIGURE 3.1: Sketch of an application scenario depicting our human in the loop visual image recognition tool for skin disease diagnosis: a patient with a skin complaint walks into a primary care clinic to be seen by a care worker. The physician then uses a smart phone to take pictures of the affected skin areas and uploads them via the mobile network to a central server. On receiving the photos and associated clinical information, a software agent in the central server will then intelligently choose an appropriate question from a pre-constructed dermatological question bank. The selected question is then sent back to the user who then sends an answer back to the agent. The Q&A is performed interactively and iteratively until the agent is sufficiently confident to output a possible diagnosis. User answers along with the visual appearance of the skin conditions are modelled together to arrive at a diagnosis.

State-of-the-art computer vision techniques for fine-grained classification of skin lesions are still far from satisfaction [136]. Therefore, a more realistic way is to utilise

human knowledge by including users in the decision-making loop. This approach boosts accuracy of such system, and also helps with the issue of trust and public alienation towards autonomous technologies [3]. We collected a large image dataset of 44 skin conditions to carry out research on human in the loop object recognition techniques for the purpose of skin conditions diagnosis. We used the crowd-sourcing tool Amazon Mechanical Turk to ask ordinary internet users a set of predefined perceptual questions and through collecting user answers to these enquiries, we obtained high-level information about the images. We assumed that users are non-medical professionals and the questions were not necessarily structured to be medically relevant. Further details about our arrangements for data collection and building the “Derm2309” dataset will be explored in the “Experiments and Results” section.

In the section 3.5 of this chapter, we will emphasise the fact that human interactions within our proposed system provide invaluable information to the skin recognition algorithms. The experiments illustrate that traditional techniques based on low-level visual features and Support Vector Machines (SVMs) can only achieve a very low recognition rate of approximately 13%, indicating the challenging nature of our medical application. However, incorporating a human in the loop approach with users answering non-medically relevant questions can boost the recognition rate to nearly 25%, achieving a performance increase of 90% over the baseline techniques. By incorporating medically relevant questions, our proposed technique can achieve a recognition rate of 48%, achieving a performance enhancement of 270% over conventional baseline techniques.

3.1 Problem Formulation

In a human in the loop classification setting, the ultimate problem is to associate an object with a meaningful class label. Hence, the problem that we aim to rectify becomes the probability estimation of an object belonging to a certain class. This is formalised in estimating a conditional probability given two inputs, where c^* is the predicted class label, c is a class label, x is image information, and S is any sequence of abstract information available from the human in the loop:

$$c^* = f(x, S) = \max_{c \in C} p(c|x, S) \quad (3.1)$$

For instance, in our human in the loop applications, users provide information in form of answers to perceptual questions that help with better classification of visual objects. We now need to examine information representation techniques and propose suitable classification algorithms.

3.2 Representation of Low-level Image Information

Image representation plays an important role in quality of any classification solution. It has been common practice [164, 165] to employ multiple visual features to represent an image. However, it is not trivial to effectively combine heterogeneous features, as we briefed in chapter 2. Common practice usually follows a weighting paradigm where similar or different weights are assigned to distinct features.

We aim to devise a random forest solution suitable for fine-grained classification in our implementation of the human in the loop recognition framework. Therefore, the method of fusing different visual features is correlated with the choice of the split function of a random tree in the forest. Given the visual feature representation F for a sample, there are many ways of defining the split function at each node. Linear classifiers [75, 110] are examples of such ways:

$$\begin{cases} w^T F + b \geq 0 & \text{go to left child} \\ \text{otherwise} & \text{go to right child} \end{cases} \quad (3.2)$$

where w is a vector of real-valued weights, and b is the bias. The above method will face difficulties when the feature dimensionality is high, as the valid search space grows rapidly and consequently selecting an appropriate split becomes more complicated. To narrow down the search space for the split function, there are dimensionality reduction methods, such as Kernel PCA [166], that construct more compact representations of each feature channel. For instance, the split function of a random tree at each node is defined as the following:

$$\begin{cases} F'_i \geq \text{thresh} & \text{go to left child} \\ \text{otherwise} & \text{go to right child} \end{cases} \quad (3.3)$$

where F' denotes the kernel PCA reduced feature.

In our proposed random forest approach, a number of feature extraction and representation techniques (e.g. SIFT [11], PHOW-HSV [167], etc.) can be deployed to form visual feature vectors of various lengths that are appropriate for describing a sample image. It is believed that a careful combination of visual features may improve the performance of a classification algorithm. We have tested a number of visual descriptors for each dataset in our experiments and have made a number of conclusions about the most discriminative features but figuring out these combinations in a brute force manner has not been the main focus of our work.

3.3 Representation of High-level User Information

We harness user-provided information in form of answers to perceptual binary questions in developing our human in the loop classification solution. These answers can be regarded as presence of tags in each image. The answers to these questions help the algorithm to classify images more efficiently. For instance, the presence of a tag “Red” in an image can be regarded as the answer to a binary question: “Is the object red?” Obviously, this is merely a simple example that can be easily answered by an algorithm purely based on visual features of an image. However, the importance of these answers (tags) become apparent when visual features fail to capture the complexity present in visually similar images of different classes.

User-provided answers can be used to build textual feature vectors with each element representing the presence of a tag. These textual vectors have length n representing n possible questions. Instead of only 0 and 1 values representing absence or presence of tags respectively, users’ answers to the binary questions can be quantified by a certainty value, i.e. guessing, definitely, and probably. These certainty values allow our framework to assign more weights to more confident answers. Therefore, each element in the vector is set as a discrete probability between 0 to 1 representing the probability of a tag belonging to an image. Any positive answer has a probability value above 0.5, and any negative one below 0.5. We set *probably* as a middle value between *definitely* and 0.5, and *guessing* as a middle value between *probably* and 0.5. Table 3.1 shows these values. Although other definitions of these numerical values are possible, we have not considered this as a focus of current studies.

TABLE 3.1: User Answers Certainties

Answer	Guessing	Probably	Definitely
Positive	0.625	0.75	1
Negative	0.375	0.25	0

The simple Graphical User Interface depicted in figure 3.2 illustrates an example of described procedure for answering perceptual questions in a medical settings, where the objective is to categorise a patient’s skin lesions into one of known conditions. The GUI represents each potential test image to the user vividly, and by providing a set of tick boxes and popup menus makes it very straightforward for an inexperienced user to answer questions or interact with the underlying framework. Ticking a box translates into a positive answer, whereas unchecked boxes are the negative responses.

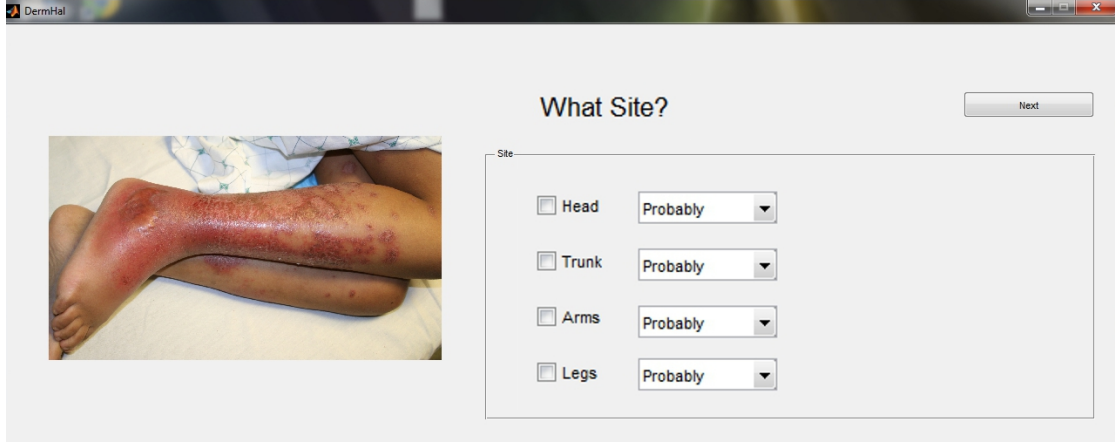


FIGURE 3.2: A skin lesion image is displayed to user. A question and its possible answers are pictured in each screen. The human operator checks the relevant answers and can assign confidence votes to quantify their responds. After answering a question, the user clicks on the “Next” button to show the next question and its relevant answers. The user will repeat this process until all questions are explored, or the maximum permissible number of questions are reached.

3.4 Human in the Loop Random Forest Classifier

The fusion of visual information x and user’s abstract source of information S is achievable both at input and output end of human in the loop classification frameworks. At the input end, fusion is performed by simply concatenating x and S together, forming a universal source of information U . The concatenated source U can be used as an input to any typical classifier including our proposed ensemble of random trees. In a probabilistic settings, this is defined as:

$$p(c|x, S) = p(c|x \cup S) = p(c|U) \quad (3.4)$$

3.4.1 Random Forest Construction

In our proposed implementation, the concatenated vectors U are exploited by a bootstrap aggregating (bagging) ensemble algorithm that follows the standard method in [72] to train random trees, and classify test samples. The widely adopted maximum information gain criteria [73], calculated based on class labels of the training images, is used in our approach as the score function to select a good split:

$$Score(split) = \Delta E = -\frac{|G_l|}{|G_n|}E(G_l) - \frac{|G_r|}{|G_n|}E(G_r) \quad (3.5)$$

where $E(G)$ is the Shannon entropy of class label distributions in the set of samples G . G_l and G_r represent the training images contained in node n 's left and right child nodes respectively. G_n is the set of training samples in node n .

3.4.2 Classification in the Random Forest

To classify a test image as some c class, it is passed through all the trained trees. It starts from the root node and traverses down to right or left nodes based on the split function, and eventually falls into one of the leaf nodes. Leaf nodes store a normalised probability distribution of the occurrence of all possible classes in the dataset. For each observation and each class, the score generated by each tree is the probability of this observation originating from this class computed as the fraction of observations of this class in a tree leaf. A common voting technique, which averages these scores over all trees in the ensemble, classifies the image.

The number of trees in the random forest has an influence on its final performance. Experiments with a few numbers are normally practised in a grid search approach to select a suitable size. Any larger size than a saturating point do not usually show a significant improvement.

3.5 Experiments and Results

We now evaluate our proposed interactive object recognition framework that is based on random forest classifiers on 6 suitable human in the loop datasets. The first 4 datasets contain medical images, and the last 2 are examples of alternative domains appropriate for interactive recognition.

3.5.1 Derm90 and Derm706 Skin Conditions Dataset

Prior to constructing the ‘‘Derm2309’’ [168], we created two smaller skin conditions datasets to perform an initial examination of our proposed human in the loop object recognition techniques for dermatology. Images of these datasets with their ground truth classification were mainly collected from: <http://www.dermis.net>. The first (Derm90) and the second (Derm706) datasets contain 90 and 706 dermatological images of 3 and 7 different skin conditions respectively.

The 3 classes of the ‘‘Derm90’’ dataset are: Discoid Eczema, Infantile Acne, and Scabies. They each contain 30 images. Allergic Vasculitis, Atopic Eczema, Bullous Pemphigoid,

Lichen Planus, Mycosis Fungoides, Squamous Cell Carcinoma, and Superficial Spreading Melanoma constitute the 7 classes of the “Derm706” dataset. They contain 72, 143, 71, 72, 83, 92, and 173 images respectively. The imbalanced distribution of class labels in this dataset is due to the varying availability of ground-truthed samples and human annotators. Amongst frequently used methods that aim to solve the problem of learning with imbalanced datasets are undersampling, oversampling, and deployment of cost-sensitive learning systems [169]. The training sets in our experiments are balanced, and we believe that the skewed distribution of class labels in the testing sets cannot have a significant impact on the final output of our classifier.

The lesions within the images were manually segmented using bounding boxes that included pixels of lesion, healthy skin and noise, such as hair. A set of 10 visual features were extracted from the entire surface of these bounding boxes, which as a whole were treated as single instances. The extracted features from individual bounding boxes were concatenated in cases where more than one box was needed to locate the affected area in the image. Table 3.2 lists the accuracy rates of these 10 conventional visual descriptors on the pilot datasets. Each of these 10 features were used to construct Chi-Squared kernels. The combination of the resulting kernels was achieved by the multiple kernel learning algorithm of [52].

Although combining multiple features is a common practice in the literature [50], it is nevertheless interesting to know which visual features will be the most discriminative for our current application. It becomes evident that PHOW-HSV [167] is the most competent feature, yielding comparable results to accuracy of all features combined together by our selected multiple kernel learning algorithm. We therefore utilise this single descriptor to represent visual information of each image in all versions of our larger skin conditions dataset. The Pyramid Histogram of Visual Words (PHOW) features [75] are a variant of dense SIFT descriptors [11], extracted at multiple scales. The colour version of PHOW extracts descriptors on the three HSV image channels and stacks them up.

TABLE 3.2: Individual Visual Features’ Accuracies on Pilot Dermatology Datasets

Feature	CPAM	GB	PHOG-180	PHOW-GREY	SIFT
Accuracy	45.97%	32.86%	27.62%	51.61%	42.54%
Feature	SSIM	GIST	PHOG-360	PHOW-HSV	DENSE-SIFT
Accuracy	34.88%	45.77%	42.34%	57.06%	46.37%

High-level user information about the images was obtained from answers to both contextual and perceptual questions such as: age of patient, history of disease in the immediate family, itchiness, contagiousness, duration of discomfort, colour, border, and shape of skin lesions. Answers to these questions were collected by fabricating medical

scenarios for a small group of 5 users from the University of Nottingham who had limited levels of clinical knowledge. Table 3.3 lists 8 questions and 36 possible answers used in the “Derm90” dataset and table 3.4 lists 13 questions and 67 possible answers used in the “Derm706” dataset.

We consulted two medical professionals from the centre of Evidence Based Dermatology at the University of Nottingham, and a dermatological reference [170] to scientifically derive these questions. The questions’ set of “Derm706” was constructed to be more comprehensive in order to compensate for the more varied skin conditions present in this dataset. Wherever specific medical terms were used, a guide image with explanations was available for users to avoid confusion.

TABLE 3.3: Derm90 Dataset Questions

Tags used as Answers to the Questions		
Site?	13. Excoriated	26. Yellow
01. Head	Lesion?	27. Orange
02. Trunk	14. Flat	28. Grey
03. Arms	15. Raised	Age?
04. Legs	16. Fluid Filled	29. Infant
Condition?	17. Broken Surface	30. Young
05. Acute	Colour?	31. Adult
06. Chronic	18. Pink	32. Old
Surface?	19. Red	Contagiousness?
07. Normal	20. Purple	33. Contagious
08. Scaly	21. Mauve	34. Non-contagious
09. Hyperkeratotic	22. Brown	Itchiness?
10. Warty	23. Black	35. Itchy
11. Crust	24. Blue	36. Non-itchy
12. Exudate	25. White	

3.5.1.1 Experiment Setup

We employed the following descriptors to form the visual feature vectors: Coloured Pattern Appearance Model (CPAM) [13], Geometric Blur (GB) [55], Global Image Descriptor (GIST) [171], Pyramid Histogram of Oriented Gradients (PHOG) and its variations [55], Scale-invariant Feature Transform (SIFT) and its variations, Pyramid Histogram of Visual Words (PHOW) and its variations [167], and Self-similarity Feature (SSIM) [55]. User provide information in form of tags are likewise used to construct textual descriptors.

The training and testing sets of “Derm90” are split by a 50:50 ratio. This translates into 15 training and 15 testing samples for each class, totalling 45 training and 45 testing images in the dataset. In the case of “Derm706” dataset, there are 30 training samples

TABLE 3.4: Derm706 Dataset Questions

Tags used as Answers to the Questions		
Age?	23. Generalised	45. Excoriation
01. Infant	Arrangement?	46. Lichenification
02. Child	24. Discrete	47. Atrophy
03. Adult	25. Coalescing	48. Papillomatous
04. Elderly	26. Disseminated	49. Warty
History?	27. Annular	50. Umbilicated
05. Personal	28. Linear	51. Shiny
06. Family	29. Grouped	Colour?
Site?	Erythema?	52. Blood
07. Face	30. Erythematous	53. Pigment
08. Scalp	31. Non-Erythematous	54. Lack of Blood/Pigment
09. Ears	Duration?	55. Others
10. Mouth/Tongue/Lips	32. Acute	56. Multicolour
11. Trunk	33. Chronic	Border?
12. Hands	Type?	57. Well-defined
13. Genitalia	34. Flat	58. Poorly Defined
14. Lower Legs	35. Raised Solid	59. Accentuated Edge
15. Feet	36. Fluid Filled	Shape?
16. Nails	37. Cyst	60. Round
Number?	38. Comedone	61. Irregular
17. Single	39. Broken Surface	62. Rectangular
18. Multiple	Surface?	63. Serpiginous
Distribution?	40. Normal	64. Dome shaped
19. Symmetrical	41. Keratinisation	65. Spherical
20. Asymmetrical	42. Scale	66. Pedunculated
21. Unilateral	43. Broken	67. Flat topped
22. Localised	44. Crust	

per class. This split ratio generates a total of 210 training and 496 testing images for the 7 classes of this dataset.

All results presented in this section are based on these features and a 5-time repeated random sub-sampling cross validation method.

3.5.1.2 Baseline Results

We initially deployed LIBSVM [172] as a baseline to examine the efficacy of our random forest solution. LIBSVM implements the “one-against-one” approach [173] for multiclass classification. We used a voting strategy in our experiments, where each binary classification was considered to be a voting. A sample is therefore designated to be in a class with the maximum number of votes. Our selected LIBSVM classifier is an RBF kernel SVM. The cost parameter of the classifier is the default value 1. The gamma parameter in the kernel function is set to the default value 1 over the number

of features. Parameter estimation using grid search with cross-validation can be also employed in the experiments.

The mean classification accuracy of LIBSVM using visual descriptors, and tuned by default parameters levels at 57.78% for the “Derm90” dataset. The LIBSVM classifier and textual features result in an accuracy of 97.13% for the same dataset. 61.09% and 96.03% are visual and textual based classification accuracies respectively for “Derm706” dataset. The combination of visual and textual features leads to mean accuracies of 97.66% for “Derm90” and 96.16% for “Derm706” datasets.

3.5.1.3 Random Forest Results

We trained our introduced random forest solution on an ensemble of 500 trees to evaluate its effectiveness in comparison to the baseline alternative. We obtain an average accuracy of 61.53% based on the extracted visual features for “Derm90”. The same test results in an average accuracy of 64.74% for “Derm706”. We also trained the same number of trees merely with our textual descriptors. The average accuracies saturate at 97.84% and 97.18% for “Derm90” and “Derm706” respectively. Random forest performs superiorly in comparison to LIBSVM in both visual and textual only tests. Once these features are combined, the classification accuracy for “Derm90” rises to 98.01%. The mean accuracy of combined descriptors levels at 97.38% in case of “Derm706”. Table 3.5 and table 3.6 summarise these results.

TABLE 3.5: Derm90 Dataset Classification Accuracies

Method	Visual Features	Tags Features	Fusion of Visual and Tags
LIBSVM	57.78%	97.13%	97.66%
RF	61.53%	97.84%	98.01%

TABLE 3.6: Derm706 Dataset Classification Accuracies

Method	Visual Features	Tags Features	Fusion of Visual and Tags
LIBSVM	61.09%	96.03%	96.16%
RF	64.74%	97.18%	97.38%

Table 3.7 reveals individual class accuracies for the “Derm90” Dataset. The mean classification accuracies of individual class labels in the “Derm706” dataset are similarly illustrated in table 3.8.

TABLE 3.7: Derm90 Dataset Individual Class Accuracies

Class	Discoïd Eczema	Infantile Acne	Scabies
Accuracy	100%	100%	93%

TABLE 3.8: Derm706 Dataset Individual Class Accuracies

Class Accuracy	Allergic Vasculitis 100%	Atopic Eczema 89.38%	Bullous Pemphigoid 100%
Class Accuracy	Squamous Cell Carcinoma 98.38%		Lichen Planus 95.23%
Class Accuracy	Superficial Spreading Melanoma 98.6%		Mycosis Fungoides 96.22%

All these empirical results highlight the usefulness of incorporating high-level knowledge in form of answers from users into our classification algorithms. The fusion results of these two pilot datasets are heavily influenced by textual descriptors due to their discriminative qualities. Even though our classification technique can achieve excellent recognition rates without visual features, we believe that computer vision plays an imperative role in reducing human labour by decreasing the number of questions users have to answer in order to arrive at a correct classification. We will further examine this concept in chapter 5.

A fully functioning system will need a question bank of hundreds, if not thousands of tags. Thus, the role computer vision plays is vital for improving efficiency. Furthermore, some images cannot be classified correctly without computer vision, even after gathering answers to all available questions. Figure 3.3 plots the recognition rates of computer vision combined with user answers and illustrates mean accuracies of the algorithm solely based on human answers to questions without visual features being involved.

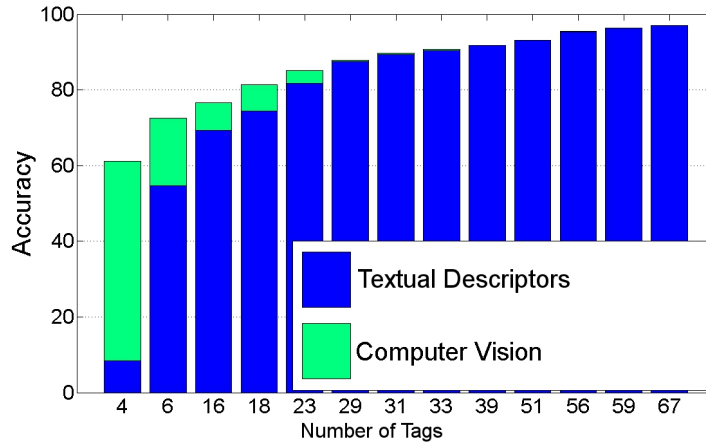


FIGURE 3.3: Computer vision plays an imperative role in reducing human labour by decreasing the number of questions required in order to achieve high accuracy results of above 80%. Textual and visual based accuracies merge approximately after asking 29 questions in Derm706 dataset. Answering all questions leads to very high mean accuracies.

3.5.2 Derm2309 Skin Conditions Dataset

We developed the challenging “Derm2309” dataset [168] over the course of 3 months. The dataset contains 2309 visually similar photographs of 44 different skin conditions. Images with their ground truth classification are all from DermIS¹. Images are not stored locally but links to original files are available in our public release of this dataset. Examples of dataset images can be found in figure 3.4.



FIGURE 3.4: Examples of skin lesion photographic images from Derm2309 dataset.
Original data courtesy of: <http://www.dermis.net>

Skin lesion images in our dataset range from different types of Eczema to various cancerous conditions, such as Superficial Spreading Melanoma. Rare conditions like Bullous Pemphigoid, as well as more common diseases like Psoriasis are also amongst the condition classes of the dataset. There are on average 52 samples per class in this dataset. At the extreme ends, Chronic Radiodermatitis has only 23 images, whilst Superficial Spreading Melanoma consists of 177 images. Table 3.9 presents the complete list of these skin conditions. As in our pilot skin conditions datasets, the imbalanced distribution of class labels in “Derm2309” dataset is due to the varying availability of ground-truthed samples and human annotators.

The training sets in our experiments are balanced, and we believe that the skewed distribution of class labels in the testing sets cannot have a significant impact on the final output of our classifier. However, there are a few cases where the number of test samples for a particular class label may remain very low, for instance in Chronic Radiodermatitis. This problem can be addressed if a leave-one-out cross validation method is applied on the dataset. We understand that this is necessary to perform in the future work in order to confidently establish the accuracy of our classifier.

There were no preprocessing steps applied on the images of the dataset. The lesions were manually segmented under supervision of a medical expert using a bounding box that

¹<http://www.dermis.net>

TABLE 3.9: Full List of Derm2309 Skin Conditions

ID	No. Samples	Condition's Name
02	37	Acrodermatitis Chronica Atrophicans Herxheimer
04	38	Acrolentiginous Melanoma
06	47	Allergic Contact Dermatitis
07	72	Allergic Vasculitis
10	31	Alopecia Areata
33	143	Atopic Eczema
28	67	Basal Cell Carcinoma
30	71	Bullous Pemphigoid
39	23	Chronic Radiodermatitis
43	59	Chronic Stationary Psoriasis Vulgaris
18	55	Dermatomyositis
19	55	Discoïd Lupus Erythematosus
22	47	Dyskeratosis Follicularis
24	36	Epidermolysis Bullosa Hereditaria
25	38	Granuloma Annulare
16	39	Hemangioma
17	35	Herpes Zoster
20	32	Ichthyosis Congenita
21	30	Incontinentia Pigmenti
23	68	Lichen Planus
42	52	Lichen Planus of the Mucosa
31	52	Lichen Sclerosus et Atrophicus
40	47	Morpheiform Basal Cell Carcinoma
32	83	Mycosis Fungoides
44	45	Nail Changes Psoriasis Vulgaris
01	44	Neurofibromatosis Generalisata
03	59	Nevocytic Nevus
05	26	Onychomycosis
08	43	Pemphigus Mucosae
09	43	Pemphigus Vulgaris
36	30	Pityriasis Rubra Pilaris Devergie
37	51	Progressive Systemic Scleroderma
38	39	Psoriasis Inversa
26	35	Pyoderma Gangrenosum
27	37	Seborrheic Keratosis
29	39	Secondary Lues
41	67	Solid-Cystic Basal Cell Carcinoma
34	92	Squamous Cell Carcinoma
35	55	Stevens-Johnson Syndrome
11	48	Subacute Cutaneous Lupus Erythematosus
12	177	Superficial Spreading Melanoma
13	38	Urticaria Pigmentosa
14	47	Verruca Vulgaris
15	37	Vitiligo

includes pixels of lesion, healthy skin, and noise such as hair. Features were extracted from the entire surface of these bounding boxes, which as a whole were concatenated and treated as single instances. Extracted features are included in the public release of our dataset [168].

High-level user information about images of this dataset is obtained from answers to questions such as: age of patient, site, number, distribution, arrangement, type, surface, colour, border, and shape of skin lesion. Medical professionals and a dermatological reference [170] were used to scientifically derive these questions. There are 37 possible answers to these questions, which can be regarded as presence of tags. Answers to these simple perceptual questions were collected from 361 “Amazon Mechanical Turk” workers. We adhered to general policies of Amazon Mechanical Turk (AMT) in the collection process, as well as seeking essential ethical approvals from the University of Nottingham for the involvement of human participants. Figure 3.5 represents a screenshot from the template used by the workers. Table 3.10 lists the type of tags used as answers to the questions that we deploy in the implementation of our solution to recognition of skin conditions. All workers’ answers are also available in the public release of our dataset.

Answer the Questions by Looking at the Image

WARNING: this HIT may contain upsetting graphical content. Worker discretion is advised!

Instructions:

- Pick the best answers to the questions
- Hover your mouse over underlined words in answers to see an example image
- Select your confidence in your picked answer
- If the answer is not clear, select your best guess, and choose the right confidence value
- The answers must describe the image, the contents of the image, or some relevant context

Questions:

Q1) What is the **age** range of person in the image?

- | | |
|---------------------------------------|--------------------------|
| <input type="checkbox"/> Infant | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Child / Teen | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Adult | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Old | Definitely (100% sure) ▼ |

Q2) What **site** of body is involved?

- | | |
|--|--------------------------|
| <input type="checkbox"/> Head | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Mouth / Tongue / Lips | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Trunk / Torso | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Arms / Hands | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Sex Organs | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Legs / Feet | Definitely (100% sure) ▼ |
| <input type="checkbox"/> Nails | Definitely (100% sure) ▼ |



FIGURE 3.5: Screenshot of “Amazon Mechanical Turk” interface used by users. Image courtesy of: <http://www.dermis.net>

3.5.2.1 Experiment Setup

Combining multiple features is a prevalent practice in implementation of computer vision and machine learning applications. It was nevertheless interesting to examine which visual features would be the most effective in recognition of skin conditions. As it was

TABLE 3.10: Derm2309 Dataset Questions

Tags used as Answers to the Questions		
Age?	Distribution?	26. Normal
01. Infant	14. Bilateral	27. Scale
02. Child/Teen	15. Unilateral	28. Broken Surface
03. Adult	16. Localised	29. Changes in thickness
04. Old	17. Generalised	Colour?
Site?	Arrangement?	30. Blood (pink/red/purple/mauve)
05. Head	18. Discrete	31. Pigment (brown/black/blue)
06. Mouth/Tongue/Lips	19. Coalescing	32. Lack of Blood/Pigment (white)
07. Trunk/Torso	20. Annular	33. Others (yellow/orange/grey)
08. Arms/Hands	21. Linear	Border?
09. Sex Organs	Type?	34. Well defined
10. Legs/Feet	22. Flat	35. Poorly defined
11. Nails	23. Raised Solid	Shape?
Number?	24. Fluid Filled	36. Round
12. Single	25. Broken Surface	37. Irregular
13. Multiple	Surface?	

established by experimenting with skin conditions pilot datasets, we decided to extract PHOW-HSV to form feature vectors of length 1024 for the “Derm2309” dataset.

PHOW is the dense SIFT [11] features applied at several resolutions. Scales at which our colour dense SIFT features were extracted are: 4, 6, 8, and 10. Each value is used as a bin size for the feature extraction function. Step in pixels of the grid, at which the dense SIFT features are extracted, was set to 5. Answers to the perceptual questions were used to construct textual features.

Training and testing images are selected randomly by a typical split for supervised learning algorithms, subject to reserving at minimum a few number of samples from every class in the testing set. This approach generates 880 training and 1429 testing images. Hence, there are 20 training samples per class, and the rest are used for testing. A standard split ratio, where 80% of images are used for training and 20% are reserved for testing, is also available in our experiment on this dataset. The 80:20 split ratio results in 1881 training and 428 testing samples.

All results presented in this section are based on a 5-time repeated random sub-sampling cross validation method.

3.5.2.2 Baseline Results

We employed LIBSVM [172] as a baseline to measure the quality of our random forest solution. Our selected LIBSVM classifier is an RBF kernel SVM as in the previous

experiment. The mean classification accuracy of LIBSVM using visual features, and tuned by default parameters levels at 13.37%. The LIBSVM classifier using tags features results in an accuracy of 14.77%. The combination of visual and tags features leads to a 16.03% accuracy. These SVM baseline results illustrate the sheer difficulty of this dataset. Adhering to a different split of data, where 80% of images are used for training and 20% for testing, yields average accuracies of 30.22%, 16.73%, and 32.61% for visual, textual, and combined descriptors respectively.

3.5.2.3 Random Forest Results

We trained our proposed random forest technique on 500 trees to evaluate its effectiveness in comparison to the SVM baseline technique. We obtain an average accuracy of 15.76% based on the extracted visual features. We also trained the same number of trees only with our tags features. The average accuracy saturates at 16.58%. Random forest performs more effectively than LIBSVM in both visual and textual only tests. More importantly, as it is clear not the visual-only nor the textual-only results are very accurate but once the features are combined, the classification accuracy rises to 25.12%. This emphasises the usefulness of incorporating high-level knowledge in form of answers from users. Table 3.11 summarises these results and table 3.12 reveals individual class accuracies. It is interesting to mention that adhering to a 80:20 split of data produces mean accuracies of 31.41%, 17.99%, and 35.98% for visual, textual, and combined descriptors respectively on this dataset.

TABLE 3.11: Derm2309 Dataset Classification Accuracies

Method	Visual Features	Tags Features	Fusion of Visual and Tags
LIBSVM	13.37%	14.77%	16.03%
RF	15.76%	16.58%	25.12%

Our tests illustrate that using tags almost always improves the individual class performances. Only in class 21 (Incontinentia Pigmenti) incorporating tags reduces the visual-only result to 10%. This is mainly due to the fact that users cannot discriminatively describe the characteristics of this particular skin condition using our predefined set of questions. However, there are classes like 42 (Lichen Planus of the Mucosa), where tags enhance visual-only results from under 19% to over 59%. Visual-only features fail dramatically on class 23 (Lichen Planus) but with the help of tags, there is an 8% improvement.

Figure 3.6 shows the accuracy of the random forest algorithm based on the number of tags randomly utilised in the solution.

TABLE 3.12: Derm2309 Dataset Individual Class Accuracies

Class	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
Accuracy	8.3	58.8	25.6	33.3	16.6	3.7	44.2	47.8	21.7	45.4	28.5
Class	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22
Accuracy	43.9	33.3	11.1	47	21	20	11.4	28.5	25	10	29.6
Class	#23	#24	#25	#26	#27	#28	#29	#30	#31	#32	#33
Accuracy	8.3	12.5	5.5	40	11.7	34	10.5	33.3	34.3	3.1	6.5
Class	#34	#35	#36	#37	#38	#39	#40	#41	#42	#43	#44
Accuracy	15.2	11.4	40	6.4	5.2	55.6	22.2	14.8	59.3	10.2	48

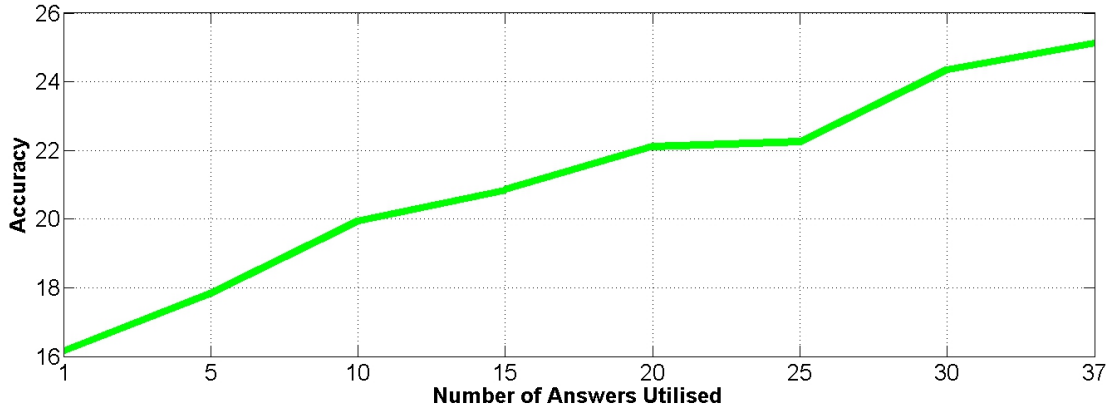


FIGURE 3.6: Number of answers incorporated, and their impact on overall accuracy of the “Derm2309” Dataset.

3.5.3 MIAS Mammographic Dataset

As a supplementary method of evaluation, we test our random forest algorithm on the MIAS database released by the Mammographic Image Analysis Society [174]. Due to popular request, the owners of the original MIAS database reduced every image to 200 micron pixel edge. They also clipped or padded every image of this dataset to 1024 pixels by 1024 pixels. Their public release contains 322 films, and the following auxiliary high-level information: character of background tissue (fatty, fatty glandular, dense glandular), class of abnormality (calcification, well-defined/circumscribed masses, spiculated masses, ill-defined masses, architectural distortion, asymmetry, normal), and severity of abnormality (benign, malignant). Figure 3.7 illustrates a few example images from this dataset.

The images in the dataset can be grouped into: Benign, Malignant, and Normal classes. There are 64 samples of the Benign, 207 samples of the Normal, and 51 samples of the Malignant class present in this dataset. As in our previous datasets, the imbalanced distribution of class labels in the dataset is due to the varying availability of ground-truthed samples and human annotators. The training sets in our experiments

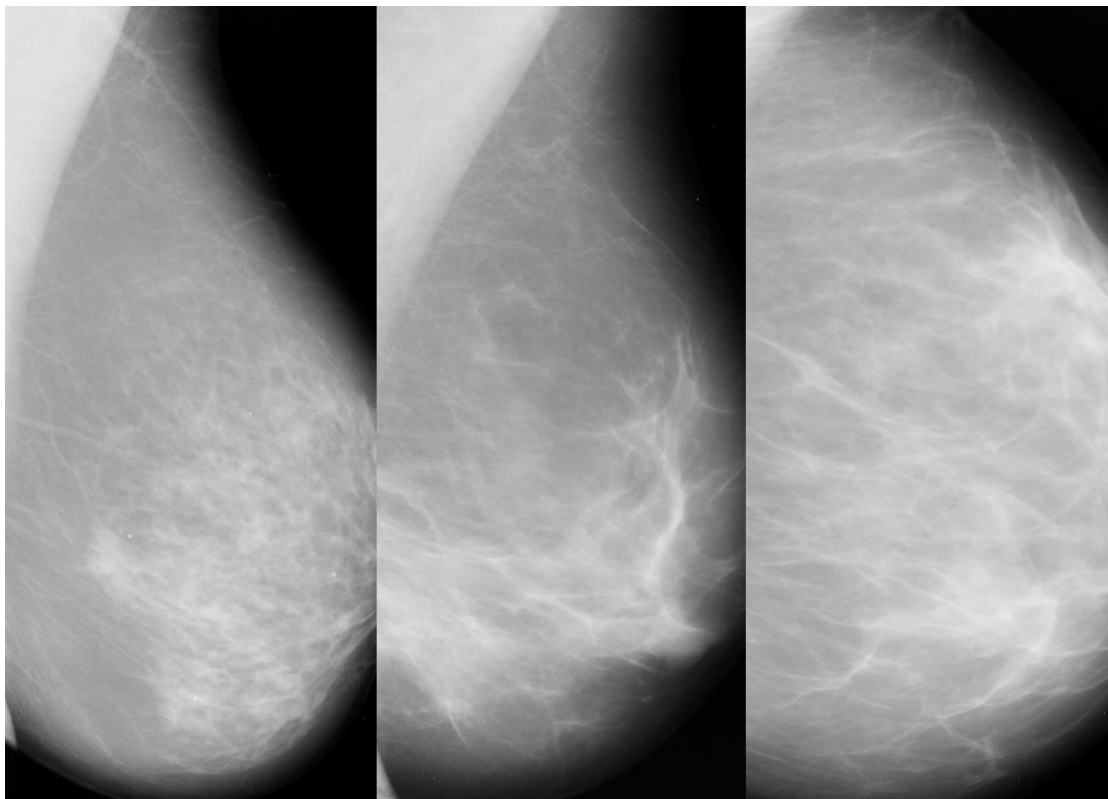


FIGURE 3.7: Examples of mammographic images from the MIAS dataset.

are balanced, and we believe that the skewed distribution of class labels in the testing sets cannot have a significant impact on the final results of our classifier.

There were no preprocessing steps applied on the images of this dataset. A number of visual features were extracted from the entire surface of the films, which as a whole were treated as single instances. The full list of extracted features are available in the “Experiment Setup” section.

A set of 10 questions based on their released high-level information are assembled together to incorporate the abstract knowledge of a user in the loop. Table 3.13 lists these questions.

TABLE 3.13: MIAS Dataset Questions

Tags used as Answers to the Questions	
Type of Abnormality?	Background Tissue?
01. Calcification	08. Fatty
02. Well-defined masses	09. Fatty-glandular
03. Needle-like masses	10. Dense-glandular
04. ill-defined masses	
05. Architectural distortion	
06. Asymmetry	
07. Normal	

3.5.3.1 Experiment Setup

We built visual feature vectors of length 5756 to represent the visual information of images in this dataset. In addition to SIFT and grey PHOG features [167] described previously, we added Grey-Level Co-occurrence Matrix (GLCM) [175], Local Binary Patterns (LBP) [167], Local Phase Quantisation (LPQ) [176], and Canny Edge Detector [177] features to the selection. To utilise our human in the loop approach, the dataset questions were exploited to build textual feature vectors of 10 dimension.

Training and testing images are randomly selected by a standard split for supervised learning algorithms, subject to reserving at minimum a number of samples from every class in the testing set. This approach generates 90 training and 232 testing images. Hence, there are 30 training samples per class and 34 testing images of the Benign, 177 testing images of the Normal, and 21 testing images of the Malignant class.

As before, all results presented in this section are based on a 5-time repeated random sub-sampling cross validation method, and the reported scores are the average of the 5 accuracies.

3.5.3.2 Baseline Results

The RBF kernel LIBSVM [172] solution produces an average classification accuracy of 14.65% using visual features. The mean classification accuracy based on textual features turns out to be approximately 88.79%. In contrast to visual features, the textual descriptors are very discriminative in this dataset. The combination of visual and textual features leads to a 88.65% accuracy. It seems that SVM baseline solution fails to exploit the discriminative power of tags available for classification of images in this dataset, as the combination recognition rates are merely as accurate as the textual results.

3.5.3.3 Random Forest Results

As in the previous cases, our random forest framework concatenates the visual and textual feature vectors to train 700 random trees. It is clear from table 3.14 that the addition of high-level human knowledge drives up the average accuracy of our proposed algorithm from 28.44% to 90.94%. It is evident that the random forest framework is outperforming the baseline's result on this dataset using the visual features. It performs almost as good as the SVM solution using the textual features. The fusion accuracy

of this framework is 90.94%, which is higher than the achieved result from the baseline solution.

In this dataset, the low-level visual features struggle mostly between the benign, and malignant classes. It is very hard to distinguish between these two classes using only visual features. However, the information provided by the human in the loop can significantly improve the accuracy of these classes. The user in the loop helps the frameworks to distinguish more confidently between the classes of this dataset. Again these results highlight the fact that the random forest framework is a simple and computationally tractable solution that produces superior, or at least comparable results to baseline solutions such as SVM on different datasets of various applications. It is clear from the results that visual features alone achieve very low recognition rates, reiterating the challenging nature of these visual recognition tasks. Human in the loop can boost accuracy rates to more acceptable levels.

TABLE 3.14: MIAS Dataset Classification Accuracies

Method	Visual Features	Tags Features	Fusion of Visual and Tags
LIBSVM	14.65%	88.79%	88.65%
RF	28.44%	88.36%	90.94%

3.5.4 Caltech-UCSD Birds 200 Dataset

Caltech-UCSD Birds 200 [178] is a dataset of 6033 images over 200 bird species from North America that cannot usually be identified by non-experts. In many cases, different bird species in this dataset are nearly visually identical. Figure 3.8 illustrates example images from this dataset.

The number of unique class labels in this dataset is 200, which corresponds to the 200 bird species. The full list of class labels is available online from the authors' released materials. This is a balanced dataset, which contains almost 30 images per class, evenly distributed in the training and testing sets.

There were no preprocessing steps applied on the images of this dataset. All samples have a bounding box that locates the bird in the image. Visual features were extracted from the entire surface of these bounding boxes, which as a whole were treated as single instances. All the extracted features are listed in the following section.

This dataset contains 25 visual questions that encompass 288 binary attributes, also referred to as tags. These attributes are extracted from <http://www.whatbird.com>, a bird field guide website. The full list of attributes can be found in their released material [178]. However, table 3.15 illustrates a sample of these tags.



FIGURE 3.8: Examples of Caltech-UCSD Birds 200 photographs.

TABLE 3.15: Caltech-UCSD Birds 200 Dataset Questions

Sample of Tags used as Answers to the Questions			
Back Colour?	Forehead Colour?	Under Tail Colour?	Crown Colour?
01. Buff	67. Grey	157. Orange	241. Blue
02. White	68. Buff	158. Yellow	242. Black
...
Back Pattern?	Head Pattern?	Underparts Colour?	Eye Colour?
16. Spotted	82. Capped	172. Grey	256. Yellow
17. Solid	83. Eyebrow	173. Yellow	257. Black
...
Belly Colour?	Leg Colour?	Upper Tail Colour?	Tail Pattern?
20. Yellow	93. White	187. Buff	270. Striped
21. Brown	94. Blue	188. Brown	271. Solid
...
Belly Pattern?	Nape Colour?	Upperparts Colour?	Throat Colour?
35. Striped	108. White	202. Buff	274. Brown
36. Solid	109. Black	203. Brown	275. Buff
...	276. Black
Bill Shape?	Primary Colour?	Wing Colour?	277. White
39. Cone	123. Brown	217. Black	278. Orange
40. Dagger	124. Grey	218. Buff	279. Grey
...	280. Yellow
Breast Colour?	Shape?	Wing Pattern?	281. Blue
48. White	138. Perching-like	232. Striped	282. Iridescent
49. Grey	139. Owl-like	233. Spotted	283. Olive
...	284. Rufous
Breast Pattern?	Size?	Wing Shape?	285. Green
63. Striped	152. Small	236. Long Wings	286. Pink
64. Solid	153. Medium	237. Broad Wings	287. Purple
...	288. Red

3.5.4.1 Experiment Setup

In our experiments, we employed 10 image features with specific parametrisation including Coloured Pattern Appearance Model (CPAM) [13], Geometric Blur (GB) [55], Global Image Descriptor (GIST) [171], Pyramid Histogram of Oriented Gradients (PHOG) and its variations [55], Scale-invariant Feature Transform (SIFT) and its variations, Pyramid Histogram of Visual Words (PHOW) and its variations [167], and Self-similarity Feature (SSIM) [55]. The textual information form a 288 dimensions tag vector that was utilised in our classification experiments.

Training and testing images are randomly selected according to the original split ratio released by the authors of this dataset. Their approach generates 3000 training and 3033 testing images. Hence, there are 15 training and approximately 15 testing samples per class present in the dataset.

All results described in this section are based on a 5-time repeated random sub-sampling cross validation method, and the reported scores are the mean of the 5 accuracies.

3.5.4.2 Baseline Results

The RBF kernel LIBSVM [172] baseline solution generates an average classification accuracy of nearly 19% using visual features. The mean classification accuracy based on high-level textual information saturates at approximately 61.92%. In comparison to visual features, the textual descriptors are more discriminative in this dataset. The fusion of visual and textual features shows a mean accuracy of 63.32%.

3.5.4.3 Random Forest Results

Our random forest framework concatenates the visual and textual feature vectors to train 1000 random trees. Table 3.16 highlights the fact that the average accuracy of our proposed algorithm improves from 20.51% to 66.32% by adding human high-level knowledge. It is also evident that the random forest framework is outperforming the baseline's result on this dataset using the visual features. The fusion accuracy of this framework surpasses the obtained result from the baseline SVM solution at 63.32%.

These results reiteratively highlight the important fact that our random forest framework is a simple and computationally tractable solution that produces superior, or at least comparable results to baseline solutions such as SVM. It is crystal clear from the results that visual descriptors alone achieve very low recognition rates, reiterating the

challenging nature of some visual recognition tasks. Human in the loop can enhance accuracy rates to more acceptable levels in challenging scenarios.

We believe that the minor drop from textual-only to fusion results on this dataset is statistically insignificant and due to the unfortunate random permutations of our training and testing samples. Furthermore, the role of textual information becomes more worthwhile, with respect to reducing the burden on users of interactive applications. We will discuss this matter in more details in the upcoming chapters.

TABLE 3.16: Caltech-UCSD Birds 200 Dataset Classification Accuracies

Method	Visual Features	Tags Features	Fusion of Visual and Tags
LIBSVM	19%	61.92%	63.32%
RF	20.51%	66.43%	66.32%

3.5.5 Ground Photograph Habitat Dataset

Torres [179] presents a geo-referenced habitat image database containing high resolution ground photographs that have been manually annotated by experts. This is the first publicly available image database specifically designed for the development of multimedia analysis techniques for ecological applications. The availability of experts' annotations in this database enables human in the loop algorithms to be employed for improved categorisation of their data. The original work from these authors presents a random forest based method for annotating an image with the habitat categories it belongs to. The authors introduce a random projection based technique for constructing a random forest classifier. Their approach is able to classify only three of the main habitat classes with a reasonable degree of confidence. Although their work has not fully examined the potential benefit of deploying a human in the loop approach, we aim to evaluate our proposed interactive method on an extended version of their adaptable dataset.

The Ground Photograph Habitat database consists of 1086 ground images with 4203 annotated polygons. There is an average of 3.85 annotations per image in this dataset, with the minimum number of distinct habitats present per image being 1 and the maximum being 6. All photographs were manually ground-truthed by an expert in Phase 1 classification. Annotation information is stored in XML files, which save the points of the polygons defining each annotation. Annotations do not overlap, since each pixel within an image uniquely belongs to one habitat. All images are geo-referenced, and they were taken with the same camera, a Sony Cybershot DSCHXvb with a 10.2 mega pixels sensor, and a 3648x2736 pixels resolution. Figure 3.9 illustrates a few example images from the dataset.

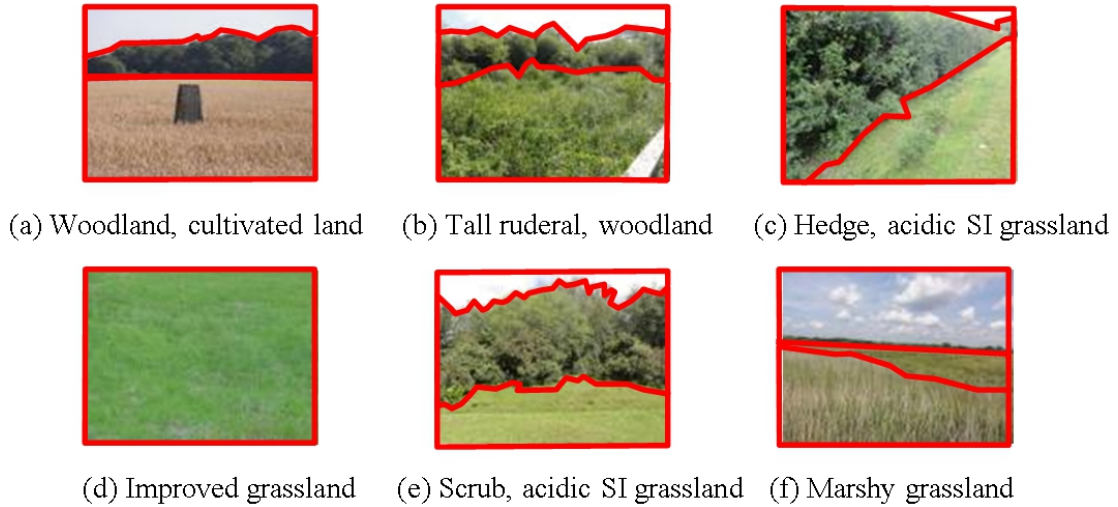


FIGURE 3.9: Examples of labelled ground-taken habitat photographs. SI stands for Semi-improved.

The photographs were taken during the months of February, July, and August in Christmas Commons, New Forest, and Titchfield Haven in the UK. Given the nature of the locations visited, mostly habitats from classes: Woodland and scrub, Grassland and marsh, Tall herb and fern, Heathland, and Miscellaneous, which includes the boundary habitats, are present in this dataset. There are 27 distinct habitats in the set.

The numbers of instances (shown inside the brackets) from each habitat available in the dataset are as follows: Woodland broad-leaved (399), Woodland mixed (242), Scrub dense (295), Scrub scattered (21), Acid grassland semi-improved (149), Neutral grassland unimproved (125), Neutral grassland semi-improved (386), Improved grassland (296), Marshy grassland (62), Poor semi-improved grassland (6), Bracken continuous (55), Bracken scattered (16), Tall ruderal (52), Dry dwarf shrub heath acid (40), Dry dwarf shrub heath basic (8), Dry heath acid grassland mosaic (88), Fern (1), Standing water (18), Cultivated arable (63), Cultivated ephemeral short perennial (1), Hedge and trees species rich (110), Hedge and trees species poor (226), Fence (231), Wall (11), Dry ditch (15), Sky (1042), Other/non-habitats (245).

The habitat dataset is a multilabel classification problem. It is a variant of the multiclass classification problem, where multiple target labels must be assigned to each instance. Formally, multilabel learning can be phrased as the problem of finding a model that maps inputs x to vectors y , rather than scalar outputs as in the ordinary classification problem. To make the dataset compatible with the scalar outputs of our multiclass random forest classification algorithms, we exploit a transformation method that maps each combination of labels present in the dataset to a unique new label. This translates to 347 unique class labels. Evaluation metrics for multilabel classification are inherently different from those used in multiclass classification, due to their inherent differences

of the classification problem. Therefore, we use the following modified metrics for the habitat dataset:

Hamming Loss (relaxed metric): the percentage of the wrong labels to the total number of labels. This is a loss function, so the optimal value is zero. $1 - loss$ equals to the accuracy.

Exact Match (strict metric): is the most strict metric, indicating the percentage of samples that have all their labels classified correctly.

There were no preprocessing steps applied on the images of this dataset. As in the previous datasets, visual features were extracted from the entire surface of these images, which as a whole were treated as single instances. The list of extracted features are available in the following section.

We transform this dataset into a compatible application of interactive recognition by introducing a set of 17 binary questions that summarises the perceptual information of images as seen by the human in the loop. The questions are listed in table 3.17. All questions were answered by inexperienced users to expand high-level information available in this dataset.

TABLE 3.17: Ground Photograph Habitat Dataset Questions

Tags used as Answers to the Questions	
Structure?	09. Reed
01. Boundary	10. Bracken or fern
02. Fence	Trees?
03. Wall	11. Trees with leaves
Landscape?	12. Trees without leaves
04. Sky	13. Trees with mixed leaves
05. Water	Field?
06. Other	14. Heath
Plants?	15. Arable land or crops
07. Bushes	16. Grass with flowers or non-uniform grass
08. Herbs	17. Uniform grass

3.5.5.1 Experiment Setup

In our solution, we employed 10 visual features to represent the habitat dataset. These descriptors are: Coloured Pattern Appearance Model (CPAM) [13], Geometric Blur (GB) [55], Global Image Descriptor (GIST) [171], Pyramid Histogram of Oriented Gradients (PHOG) and its variations [55], Scale-invariant Feature Transform (SIFT) and its variations, Pyramid Histogram of Visual Words (PHOW) and its variations [167],

and Self-similarity Feature (SSIM) [55]. These visual descriptors result in formation of 8976-dimensional feature vectors. Textual features were constructed based on the 17 questions, we described in the last section.

Training and testing images are randomly selected and split into a common ratio for supervised learning algorithms, subject to reserving at minimum a few number of samples from every class in the testing set. This approach generates 657 training and 429 testing images. Our selected split ratio also ensures that all the 347 unique classes of the dataset are present in the training set and are available to be used by the learning algorithm.

All results presented in this section are based on a 5-time repeated random sub-sampling cross validation method, and the reported scores are the average of the 5 accuracies.

3.5.5.2 Baseline Results

Our selected LIBSVM classifier [172] is an RBF kernel support vector machine, similarly tuned to our setups in the previous datasets. The SVM solution based on visual features achieves relaxed and strict accuracies of 38.91% and 3.03% respectively. Results based on textual features are 44.22% and 5.94%. The combination of visual and textual features in the baseline solution obtains accuracies of 50.32% for relaxed and 10.78% for strict metrics.

3.5.5.3 Random Forest Results

The low-level visual features in both the SVM baseline approach and our random forest solution particularly struggle to distinguish between semi-improved, and unimproved grassland classes. These classes are even subjective for human surveyors. Additionally, broad-leaved trees can be part of both the Woodland habitat, which is composed of broad-leaved trees, and the Mixed Woodland habitat, which is itself composed of broad-leaved trees and coniferous trees. This similarity in classes explains the reason why the low-level features may struggle to classify these habitats. In these cases, the value of human in the loop additional information becomes very clear. A few simple perceptual binary answers from a user can help the framework to make the right decision.

Our random forest framework concatenates the visual and textual features together, and trains 700 random trees. Table 3.18 summarises the accuracies of our random forest framework using different types of features. The results of both evaluation metrics we described previously are presented. As it is clear, the combination of low-level

visual features with high-level knowledge of users increases the average accuracy of our algorithm to 57.22%. This enhancement is true with both metrics.

TABLE 3.18: Ground Photograph Habitat Dataset Classification Accuracies

Baseline SVM			
Metric	Visual Features	Tags Features	Fusion of Visual and Tags
Strict	3.03%	5.94%	10.78%
Relaxed	39.91%	44.22%	50.32%
Random Forest Framework			
Metric	Visual Features	Tags Features	Fusion of Visual and Tags
Strict	13.75%	11.65%	17.94%
Relaxed	53.24%	52.81%	57.22%

It is clear that our random forest framework is outperforming the baseline SVM solution in every aspect of the evaluation. The random forest framework is a simple and computationally tractable solution that produces interesting results on this dataset using both strict and relaxed metrics.

3.6 Conclusion

We have come to believe that a human in the loop approach that combines high-level cognitive information with traditional low-level visual features offers the possibility of developing practically useful machine vision technologies. This is particularly true whilst the pursuit for fully automatic solutions continues. Human in the loop enables realisation of applications that are believed to be impractical, or too critical to be left for computers to administer single-handedly.

Our random forest framework is a reliable solution that can be improved to produce practical results for a range of different applications. We strongly believe that by working closely with dermatologists for instance, our work can be improved and expanded to practical levels suitable for health care providers across the world. An enhanced solution can be installed on smart mobile phones or tablets, and used by physicians to improve patients' quality of lives in both developed and developing countries where access to health services is scarce.

In this chapter, we proposed a simple but powerful method to utilise and quantify user answers to simple perceptual questions in a systematic way that can be incorporated into our introduced framework of fusing heterogeneous sources of data. Furthermore, we formulated a method based on random forest technology that combines visual features of images with their relevant user abstract information to achieve promising recognition

rates. Our proposed framework has the capability to be used in an interactive fashion. Once in testing mode, users of our system can answer an application's questions as little as they desire, and still receive a prediction based on available information. Certainly, more answers means more information, which in turn leads to more solid classification accuracies.

In the following chapters, we will discuss more intricate methodologies for fusing various information sources, and will examine potential solutions for achieving the most effective fusion based on available information sources. We will also address the issues centred around user involvement in human in the loop frameworks, and possible remedies to reducing their burden, whilst still being able to capture their useful abstract knowledge.

Chapter 4

Generative Object Recognition with Humans in the Loop

Amongst major human in the loop technical complications is the problem of information source fusion. In the “Understanding Visual Content by Human in the Loop” section of chapter 2, we defined the term “fusion” in the context of our work as a data fusion process, where the sources of data are image’s low-level and users’ high-level information. In the preceding chapter, we introduced a discriminative random forest approach that exploits user’s abstract knowledge in a framework, where both visual and textual sources of information are considered to be analogous. Consequently, they are concatenated to form one single source of information prior to being fed to the random forest classifier. Contrary to these input level fusion algorithms, there are methodologies that model visual and textual descriptors separately to conclude a final prediction.

The fusion of low-level visual information and high-level human knowledge has been previously achieved by the pioneering framework of [130], which assigns equal weights to all available sources of information. Their general framework is capable of employing almost any off-the-shelf multiclass object recognition algorithm. They further illustrate that incorporating models of stochastic user responses leads to better reliability in comparison to deterministic field guides generated by experts. Their evaluation results demonstrate that utilising user input drives up recognition accuracy to levels that are sufficient for practical applications, whilst at the same time, computer vision reduces the amount of human interaction required.

Nevertheless, it is reasonable to assume that these different origins of information are not always equally reliable or discriminative. For instance, in a classification setting, outputs from a typical visual classifier may not be as accurate as the outputs from a classifier trained on user provided information. The opposite may also be valid, where

user knowledge is more misleading, vague and noisy than the visual information of objects in an image. Therefore, it is sensible to investigate either methodologies that are capable of selecting the most informative and reliable source, or algorithms that are competent enough to intelligently assign appropriate weights to each and every source of information available.

In this chapter, we mainly aim to introduce solutions to the aforementioned problem of information source fusion. Nevertheless, we will likewise explore appropriate classification methods of visual and textual information. Our proposed innovative algorithms are:

1. A modified naive Bayes algorithm that adaptively selects an individual classifier's output or combines more to produce a definite answer.
2. A neural network based algorithm which feeds the outputs of classifiers to a 4-layer feedforward network to generate a final output.
3. A novel generative model based on random naive Bayes classifiers to capture and analyse abstract information from users.

Our proposed methods intelligently combine available sources of information in order to enhance classification performance for difficult visual recognition tasks. To illustrate the efficacy of our proposed approaches over traditional fusion techniques [130], we present experimental results on a variety of computer vision datasets suitable for human in the loop object recognition.

4.1 Problem Formulation

As before, the problem that we aim to solve is to find the probability of an object belonging to a certain class. This is formalised in estimating a conditional probability $p(c|x, S)$ given two variables, where c is class, x is image information, and S is any sequence of abstract information available from the human in the loop. The fusion of x and S is achievable both at the input or the output end of classification algorithms as illustrated in figure 4.1.

1) At input level, as we detailed in chapter 3, fusion is performed by simply concatenating x and S together, and forming a universal source of information U . The concatenated source U can be used as an input to any typical classifier. In a probabilistic setting, this is defined as:

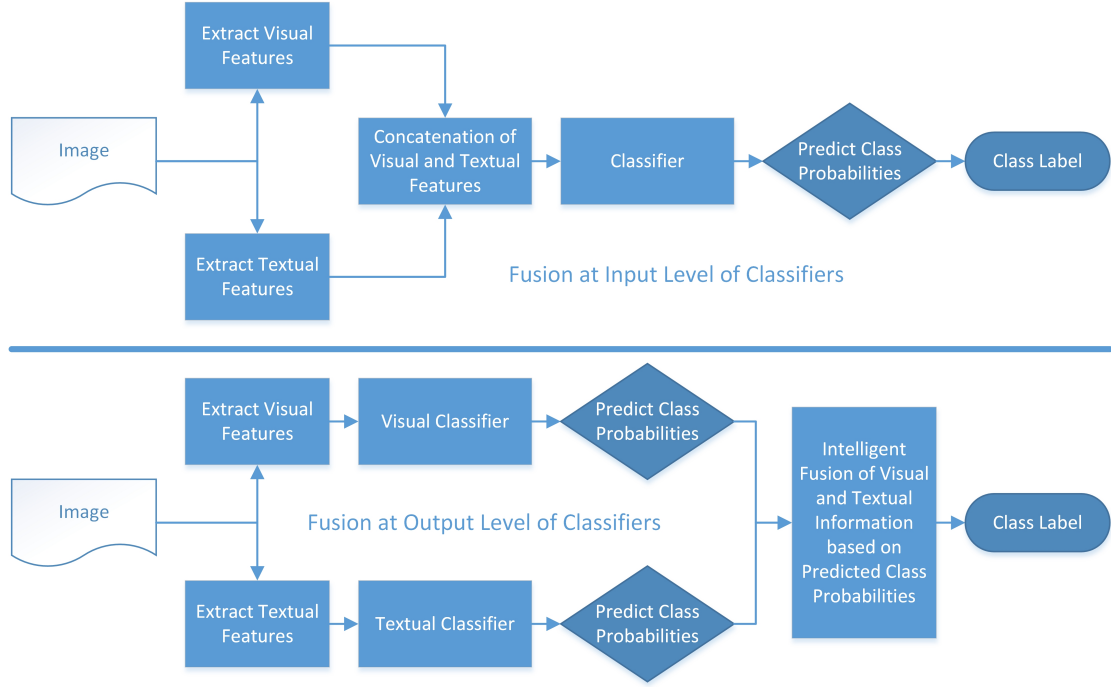


FIGURE 4.1: Fusion Frameworks at Input and Output Levels

$$p(c|x, S) = p(c|x||S) = p(c|U) \quad (4.1)$$

where $||$ is our selected notation for illustrating mathematical concatenation.

2) Fusion at output level in contrast to the previous case combines the output of classifiers independently trained on the two sources of information to produce an overall output in form of:

$$p(c|x, S) = f\left(p(c|x), p(c|S)\right) \quad (4.2)$$

where f is the fusion function, and $p(c|x), p(c|S)$ are the posterior probabilities obtained from separate classifiers trained on visual and textual information respectively. Depending on the form of f , we can design a variety of fusion models.

In spite of simplicity in implementation of fusion methods at input end of classifiers, concatenation at this level may not fully appreciate the discriminative capacity of each information source. Alternatively, we aim to propose a solution that learns separate models for each source of information available. Thus, we employ typical classifiers commonly used in the relevant literature to classify individual information sources and present two intelligent strategies to implement the fusion function f of equation (4.2).

4.2 Intelligent Information Source Fusion

It is generally assumed that class posterior outputs from probabilistic classifiers can be considered as certainty measures, if the training and testing sets are randomly selected from the same distribution. This is a reasonable assumption for the learning problems that we are targeting to solve. Hence, this hypothesis leads us to formulation of the following two strategies.

4.2.1 A Modified Naive Bayes Information Fusion Algorithm

The information fusion in equation (4.2) becomes the classic naive Bayes classifier when the two sources of information are assumed independent. This can be formally expressed by defining the fusion function f as:

$$f(x, S) = xS \rightarrow f(p(c|x), p(c|S)) = p(c|x)p(c|S) \quad (4.3)$$

where x and S are the image and user provided information respectively.

We have observed through experiments that for certain classes, decisions made on one information source can be more reliable than the other. It is therefore reasonable to speculate that if the probability of a class estimated from one source is too insignificant, then that source is very likely to be unreliable for predicting the class label. Based on this rationale, we present a modified algorithm of naive Bayes information fusion.

Algorithm 1 Modified Naive Bayes Information Fusion $p(c|x, S)$

Require: image information: x , user information: S , class labels: C

```

for all Samples do
  if  $p(c|x) < \Theta_x[c]$  then
     $p(c|x, S) \propto \frac{p(c|S)}{p(c)}$ 
  else if  $p(c|S) < \Theta_S[c]$  then
     $p(c|x, S) \propto \frac{p(c|x)}{p(c)}$ 
  else
     $p(c|x, S) \propto \frac{p(c|x)p(c|S)}{p(c)}$ 
  end if
end for
return  $\arg \max_c p(c|x, S)$ 

```

The preceding algorithm 1 is very straightforward. For each class c , we estimate a threshold $\theta_x[c]$ for visual image information, and a separate threshold $\theta_S[c]$ for user's source of information. If the probability of a class estimated from one source is smaller than its threshold, then only the probability estimated based on the other source is

used to predict the class. The original naive Bayes algorithm [130] is utilised when the probabilities of a class estimated from both sources are greater than their respective thresholds. If the estimated probability is smaller than both thresholds, it does not matter which classifier is employed. We will illustrate in the experimental section that for certain applications this modification can significantly improve accuracy over the classic naive Bayes classifier, which is indeed a special case of our intelligent fusion algorithm.

The optimal thresholds for each class and every source of information are estimated by a grid search approach that exhaustively examines a range of possible values. The selected threshold for each class is a value that leads to the finest classification performance over the training dataset of samples with known class labels $T = \{(G_i, C_i) : i \in [n]\}$. This is achieved by minimising the empirical risk:

$$L(\Theta; T) = \frac{1}{n} \sum_{i=1}^n l(C_i, f(G_i; \Theta)) \quad (4.4)$$

where Θ is a set of thresholds to be learned, and l is measure of error between groundtruth C_i and predicted $f(G_i; \Theta)$ labels. Threshold values in practice filter out uncertain predictions from deployed classifiers in our fusion framework. Algorithm 2 summarises our method in selecting suitable thresholds.

Algorithm 2 A Grid Search Approach for Optimal Threshold Selection

Require: class label: c , matrix of posterior probabilities: P

Step 0: Generate a discrete set of possible thresholds: $\Theta = \{0, 0.1, 0.2, \dots, 1\}$

Step 1: Create an empty set to store scores of each threshold: $Scores = \emptyset$

while there exist unexamined $\theta \in \Theta$ **do**

Step 2: Create an empty set to store predicted labels: $K = \emptyset$

for all samples with true label c **do**

Step 3: Find vector of posterior probabilities: $Vec = P_{j,:}$

Step 4: Find probability of the most probable class: $p = \arg \max(Vec)$

if $p > \theta$ **then**

Accept label: $K = K \cup label$

end if

end for

Step 5: Calculate F_1 measure obtained by threshold θ

Step 6: Store calculated measure: $Scores = Scores \cup measure$

end while

return $\arg \max_{\theta}(Scores)$

4.2.2 Neural Network Fusion Algorithm

The method discussed previously is a greedy approach. It follows the problem solving heuristic of making the locally optimal choice in selecting suitable weights for every predicted class label with the aim of finding a global optimum. The weights are in essence the calculated thresholds for outputs from classifiers. However, greedy algorithms usually fail to find the globally optimal solution. Our proposed greedy approach does not operate holistically on all class labels. It examines each class label at a time, and hence can make commitments to certain choices too early, which prevents it from finding the best overall solution afterwards. It may even produce the unique worst possible solution.

In an attempt to find the global optimum, we propose a pattern recognition solution that trains a supervised neural network to produce desired outputs in response to sample inputs. More specifically, we intend to deploy a feedforward backpropagation network [180]. Our selected choice of network training function is a scaled conjugate gradient backpropagation approach [181] that updates weight and bias values according to the scaled conjugate gradient method.

The architecture of the neural network we aim to use has 4 layers, as depicted in figure 4.2. The input layer has $2n$ input units. i_k , where $k = \{1, 2, \dots, n\}$, is the predicted probability of class k based on the image's visual source of information using a standard classifier. Some of these classifiers are described in section 4.3 of this chapter. Similarly, S_k , where $k = \{1, 2, \dots, n\}$, is the predicted probability of class k based on the user provided textual source of information. We employ two hidden layers in our implementation and their number of units is determined experimentally. The output layer has n units, each corresponding to one of the class labels. In preparing the “desired” output for an input training sample, we set the corresponding unit's desired output to 1 and the rest to 0. For instance:

$$L_1 = 0, L_2 = 0, \dots, L_{k-1} = 0, L_k = 1, L_{k+1} = 0, \dots, L_n = 0 \quad (4.5)$$

is the desired output corresponding to a training sample belonging to class k . Once the network is trained, the final decision about the class label is made based on the following:

$$c^* = \arg \max_k L_k \quad (4.6)$$

We further need to calculate a network performance that leads to good classification. Thus, we suggest to minimise a cross-entropy term [182] given targets, outputs,

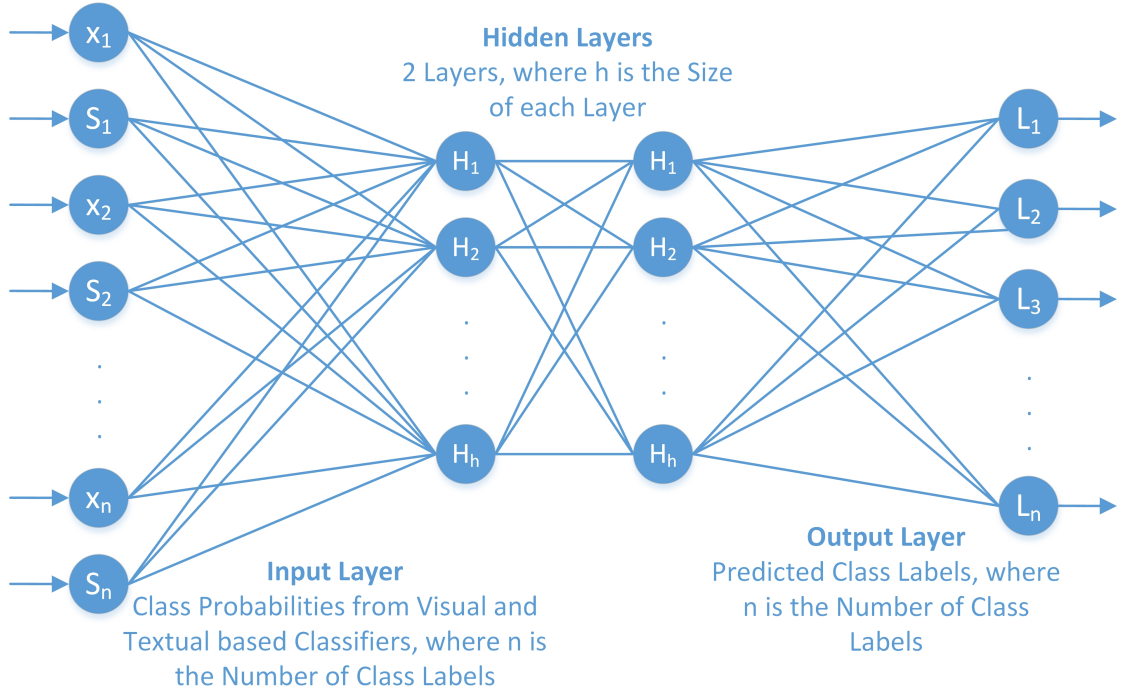


FIGURE 4.2: Neural Network Layout

performance weights, and optional parameters. Our selected entropy term is therefore defined as:

$$H(C, \hat{C}) = - \sum_i C(i) \log \hat{C}(i) \quad (4.7)$$

where $H(C, \hat{C})$ is the computed cross-entropy of true and predicted class labels, which heavily penalises outputs that are extremely inaccurate, with very little penalty for fairly correct classifications.

4.3 Classifiers in Fusion Frameworks

The classification of available information from images and users involved in human in the loop vision applications is usually performed by a number of commonplace techniques. Fusion at input level for instance can be carried out by an SVM solution or an ensemble approach like our random forest method in chapter 3 that concatenates visual and textual descriptors. At output level, it is plausible to apply a combination of similar or different classifiers on available sources separately. Table 4.1 summarises these common settings in human in the loop frameworks. Please note that RF and RNB stand for Random Forest and Random Naive Bayes respectively in this table, and $||$ is the selected notation for illustrating mathematical concatenation of descriptors.

TABLE 4.1: Classifiers Settings in Human in the Loop Frameworks

Feature	Visual	Textual	Visual Textual
Input Level	N/A	N/A	SVM, RF
Output Level	SVM, RF	SVM, RF, RNB	N/A

In the next section, we introduce our proposed generative classifier suitable for harnessing user knowledge, and our selected discriminate method for classification of low-level visual information in human in the loop applications. In the experiment section of this chapter, we will compare the results of these classifiers to their common alternatives.

4.3.1 A Generative Model for High-level User Information

We build an innovative random naive Bayes model to estimate the posterior probability of $p(c|S)$. This allows us to classify a user’s high-level information effectively. Our approach learns the class-conditional density $p(S|c)$ by a generative model that estimates user answers S for each possible class label c .

4.3.1.1 Presentation of High-level Information

As in our previous works [157, 158, 160, 161] described in chapter 3, we collect high-level information about images in the form of answers to perceptual questions. These answers can be regarded as presence of tags in each image. The importance of these tags become apparent when visual features fail to capture the complexity present in visually similar images.

Suppose there are T possible tags in our problem. Let $t \in \{1, \dots, T\}$ be an array of indices to those T tags, and let $S = \{s_1, \dots, s_{t-1}, s_t\}$ be a set of user answers about the presence of such tags in an image. Then an image can be represented as a vector of tags.

To deal with user reasoning that is approximate rather than exact, we quantify presence of tags in an image by a certainty value that describes user confidence in their response. This is in contrast to the traditional binary approach, where tags’ random variables take on only true or false values. Our answered tags random variable s_t have a discrete truth value that ranges in an interval between 0 and 1, corresponding to their chance of presence in an image. These certainty values $v \in \{1, \dots, V\}$ allow the model to assign more weight to more confident answers. Any positive answer has a probability value above 0.5, and any negative one is below 0.5. We set analogous terms for these certainty values like *probably* as a middle value between *definitely* and 0.5, and *guessing* as a middle value between *probably* and 0.5. Table 4.2 shows these certainty values, where

$V = 6$. Although other definitions of these numerical values are possible, we have not considered this as a focus of current studies and retained original values set in the previous chapter.

TABLE 4.2: User Answers Certainties

Answer	Guessing	Probably	Definitely
Positive	0.625	0.75	1
Negative	0.375	0.25	0

4.3.1.2 Modelling User Answers

Our proposed generative model for estimating user provided high-level information needs the class conditional density $p(S|c)$ to be specified. We make the assumption that questions are answered by the user independently given the class, and any randomness in their response is not image dependant:

$$p(S|c) = \prod_t^T p(s_t|c) \quad (4.8)$$

We estimate $p(s_t|c)$ separately for each value of c , thus we only solve C separate density estimation problems. An expedient strategy to avoid the problem of exponential explosion is to naively assume that the parameters of such conditional distribution are independent. Since $s_t \in \{1, \dots, K\}$, the multinomial class-conditional density for each $p(s_t|c)$ is defined as:

$$p(s_t|c, \theta_c) = \prod_{k=1}^K \theta_{ck}^{I(s_t=k)} \quad (4.9)$$

Similarly, we fit a multivariate multinomial distribution to our discrete training vector:

$$p(S|c, \theta) = \prod_{t=1}^T \prod_{k=1}^K \theta_{tck}^{I(s_t=k)} \quad (4.10)$$

where $\theta_{tck} = p(s_t = k|c)$ is the probability of observing the t^{th} tag being k given that the class label is c .

4.3.1.3 Ensemble of Random Naive Bayes Classifiers

Frequently, an ensemble of models performs superiorly in contrast to any individual model. To take advantage of such increased stability, we propose to employ an ensemble averaging process, where we train a group of random naive Bayes Classifiers. Each of these individual classifiers may overfit the training data. However, the combination of these classifiers usually results in a new network that alleviates the problem of sensitivity to random noise in the training set.

In an ensemble learning method, injection of randomisation leads to decorrelation between the individual classifiers, and improved generalisation [71, 72]. We introduce randomisation in our ensemble model by two mainly used means of: i) random input selection, and ii) random feature selection. These procedures of injecting randomisation help us achieve higher robustness with respect to presence of noisy data in user responses.

Firstly, given our collected training set S of size n , a bootstrap aggregating technique is used to generate m new training sets S_i , each with the same size n , by sampling from S uniformly and with replacement. By sampling with replacement, some observations may be repeated in each S_i . Every set S_i is expected to have a fraction of the unique examples of S , the rest being duplicates. The B random naive Bayes models are subsequently fitted using the above m bootstrap samples.

Secondly, we randomly select F features out of user responses' feature pool S for each random naive Bayes classifier. The probability distribution $p(s_f|c)$ of each feature is therefore modelled for each class c . The probability of a sample s belonging to the class c can eventually be defined as:

$$p(s|c) \sim \sum_{b=1}^B \prod_{f=1}^F p^b(s_f|c) \quad (4.11)$$

where we combine B randomly trained naive Bayes classifiers, each using a subset of available features $F \leq |T|$, by a voting scheme for the final classification. The scheme selects the class label with the highest number of votes as the predicted label.

4.3.2 A Discriminative Classifier for Image Low-Level Information

In our proposed implementation, the visual model of fusion equation (4.2) is designed to directly learn a function that computes the class posterior $p(c|x)$. This is therefore defined as a discriminative model that discriminates between different classes given the feature input.

It is sensible to reiterate that image representation plays an important role in the quality of any visual classification solution. It is believed that a careful combination of visual descriptors may improve performance of the classification algorithm but this is not the main focus of our current work. Instead, we aim to use a selection of well-known visual words with specific parametrisation to form visual feature vectors suitable for our classifiers.

Our main selected discriminative model of visual information is based on a bootstrap aggregating ensemble algorithm that follows the standard method in [72] to train random trees, and classify test samples. The widely adopted information gain criteria [73], calculated based on class labels of the training images, is used as the score function to select a good split:

$$Score(split) = \Delta E = -\frac{|G_l|}{|G_n|}E(G_l) - \frac{|G_r|}{|G_n|}E(G_r) \quad (4.12)$$

where $E(G)$ is the Shannon entropy of class labels distributions in the set of samples G . G_l and G_r represent the training images contained in node n 's left and right child nodes respectively. G_n is the set of training samples in node n . Leaf nodes store a normalised probability distribution of the occurrence of all possible classes in the dataset.

4.4 Experiments and Results

We now illustrate the effectiveness of our proposed intelligent fusion techniques suitable for incorporating human abstract knowledge in the decision making loop of visual object recognition tasks. We have tested our solution on 4 datasets appropriate for evaluating human in the loop applications. There are 2 examples from medical settings, and 2 applications of fine-grained visual classification.

The combination results at output level are based on two discriminative methods for visual and textual features: i) an RBF kernel Support Vector Machine tuned similarly to our experiments' setup in the previous chapter, and ii) an ensemble of 1000 bagged decision trees that we described previously in 4.3.2. We also have two generative approaches suitable only for textual information from users: i) method of [130], and ii) our introduced 1000 random naive Bayes solution in 4.3.1. All results presented are based on a 5-time repeated random sub-sampling cross validation method. It may be necessary to clarify that the fusion accuracies at input level presented for random forest and SVM methods are the results of concatenating visual and textual feature vectors, similar to our experiments in the previous chapter.

4.4.1 Derm2309 Skin Conditions Dataset

This previously introduced dataset [168] contains images of skin conditions from 44 different diseases. Details of the dataset were fully described in the preceding chapter. Table 4.3 illustrates classification accuracies of this dataset using different features and methods.

4.4.1.1 Visual Results

The SVM solution has a mean classification accuracy of 13.37%. The random forest technique results in an average accuracy of 15.76%. These classifiers are both fed with the same type of visual features available from the public release of this dataset.

4.4.1.2 Textual Results

The SVM classifier using textual features results in an accuracy of 14.77%. A random forest trained with the textual features has an average accuracy of 16.58%. The learned model of [130] based on the multinomial distribution results in a mean accuracy of 18.4% on this dataset. Our proposed random naive Bayes method has an average accuracy of 21.02%, a better performance than the SVM, and random forest approach. This reiterates the effectiveness of our proposed solution to modelling high-level user information in this dataset.

TABLE 4.3: Mean Accuracy of Classification Algorithms on Derm2309 Dataset

Information Source	Classification Technique	Mean Accuracy
Visual Based	SVM	13.37%
	Random Forest	15.76%
Textual Based	SVM	14.77%
	Random Forest	16.58%
	Naive Bayes [130]	18.4%
	Random Naive Bayes	21.02%

4.4.1.3 Combination Results

The most effective fusion solution for this dataset is again proved to be the neural network approach based on random forest visual and random naive Bayes textual classification techniques. It is quite interesting to note that in table 4.4 a simple concatenation of visual and textual features and utilising a SVM algorithm yields a classification accuracy of merely 16.03%. However, the deployment of neural network

approach based on the same visual and textual features results in a mean accuracy of 34.81%. This is an improvement of 18.78%. Baseline’s combination [130] accuracy result of this dataset saturates at 22.39% using a visual SVM and a textual naive Bayes classification technique.

Once visual and textual features are combined at input level using the random forest technique, the classification accuracy rises to 25.12%. These improvements show the usefulness of high-level knowledge in shape of answers from users. Our most effective version of the greedy solution outperforms the Bayesian baseline combination in [130] by approximately 8%. Our neural network approach performs better than other fusion techniques with a 34.81% mean accuracy. These results also illustrate the importance of source selection, and our proposed intelligent fusion methods.

TABLE 4.4: Mean Accuracy of Fusion Algorithms on Derm2309 Dataset

Fusion	Classifier	Concatenation		
Input Level	SVM	16.03%		
	RF	25.12%		
Fusion	Classifier	Equal Weight	Greedy Alg.	Neural Net.
Output Level	SVM+SVM	18.7%	20.98%	23.9%
	RF+RF	25.58%	28.7%	31.74%
	RF+RNB	26.08%	30.95%	34.81%

4.4.2 MIAS Mammographic Dataset

As another potential human in the loop medical application, we test our algorithms on the aforementioned MIAS database released by the Mammographic Image Analysis Society [174]. Full details of our experiment’s setup for this dataset can be found in the previous chapter.

4.4.2.1 Visual Results

We build visual feature vectors of 5756 dimension to represent the visual information of images in this dataset. In addition to the SIFT and PHOG features, we add Grey-Level Co-occurrence Matrix (GLCM) [175], Local Binary Patterns (LBP) [167], Local Phase Quantisation (LPQ) [176], and Canny Edge Detector [177] descriptors to the selection.

In this dataset, the low-level visual features struggle mostly between the benign, and malignant classes. It seems that it is very hard to distinguish between these two classes using only visual features. The random forest classifier is outperforming the SVM baseline result on this dataset using the same visual features, as seen in table 4.5.

4.4.2.2 Textual Results

The answer to the binary questions can be used to build user response pairs suitable for the Bayesian method described in [130]. Each user response s contains an answer, and a confidence value. As before, to utilise the random forest classifier, the questions are exploited to build a textual feature vectors of 10 dimension. Random forest performs almost as effective as the naive Bayes model in [130] using merely textual features. Our proposed random naive Bayes method produces comparable mean accuracies to other accurate approaches on this dataset.

TABLE 4.5: Mean Accuracy of Classification Algorithms on MIAS Dataset

Information Source	Classification Technique	Mean Accuracy
Visual Based	SVM	14.65%
	Random Forest	28.44%
Textual Based	SVM	88.79%
	Random Forest	88.36%
	Naive Bayes [130]	89.65%
	Random Naive Bayes	89.65%

4.4.2.3 Combination Results

Table 4.6 summarises the mean accuracies of various techniques using different sources of information. The combination of low-level visual features with high-level knowledge of human in the loop leads to an average accuracy of 89.65% using the Bayesian method [130] of equal weight fusion. Due to the discriminative nature of questions in this dataset, the mean accuracy of the framework using textual features is very high. However, the Bayesian fusion method fails to exploit the information contained in the visual features to improve the combination accuracy. The fusion accuracy is only as precise as the tags' results on this dataset using the Bayesian or alternative techniques. Our introduced greedy algorithm of source selection can improve the performance slightly but our neural network approach exploits the visual information more efficiently, and enhances the fusion accuracy to 94.81% from the baseline result of 89.65% produced by the Bayesian framework.

At input level, the random forest framework concatenates the visual and textual feature vectors. The addition of human high-level knowledge drives up the mean accuracy of the algorithm from 28.44% based on visual descriptors to a staggering 90.94%. It is clear from the results that visual features alone achieve very low recognition rates, reiterating the challenging nature of these visual tasks. Nevertheless, human in the loop knowledge

can boost recognition rates to more acceptable levels. Our proposed fusion techniques produce enhanced results in comparison to most accurate solutions on this dataset.

TABLE 4.6: Mean Accuracy of Fusion Algorithms on MIAS Dataset

Fusion	Classifier	Concatenation		
Input Level	SVM	88.65%		
	RF	90.94%		
Fusion	Classifier	Equal Weight	Greedy Alg.	Neural Net.
Output Level	SVM+SVM	88.45%	89.25%	89.25%
	RF+RF	89.23%	90.4%	91.23%
	RF+RNB	90.9%	90.08%	94.81%

4.4.3 Caltech-UCSD Birds 200 Dataset

CUB-200 [178] is the familiar dataset from our previous chapter. It includes 6033 images over 200 bird species, such as Myrtle Warblers, Pomarine Jaegers, and Black-footed Albatrosses. Details of the experiment’s setup were previously described.

4.4.3.1 Visual Results

The computer vision algorithm in [130] is based on Andrea Vedaldi’s publicly available source code [55], which combines vector-quantised geometric blur and colour/grey SIFT features using spatial pyramids, multiple kernel learning, and per-class 1-vs-all SVMs. The authors also add features based on full image colour histograms and vector-quantised colour histograms. They use a validation set to tune parameters for the visual classification $p(c|x)$. For comparative purposes, we also test this dataset using our proposed discriminative random forest solution with the same visual features.

The main advantage of employing computer vision on this dataset is to reduce human labour by minimising the number of questions user has to answer, or in other words the number of tags needed to improve the quality of classification predictions. Computer vision is more effective at reducing the average amount of time than reducing the time spent on the most challenging images. It is clear from the results in table 4.7 that random forest outperforms the SVM algorithm on this dataset.

4.4.3.2 Textual Results

A deterministic user with precise responses is assumed to achieve perfect classification accuracy on this dataset within the first few rounds of answering questions. However,

this assumption is not realistic, since subjective answers by user are common and unavoidable. Stochastic user responses increase the number of questions necessary to achieve a certain accuracy level. It is important to note that some images in this dataset can never be classified correctly without computer vision, and solely by utilising user answers.

We represent the classification accuracy results when merely user responses are incorporated without any computer vision involved in the process in this section. The learned model of [130] based on multinomial distribution results in a mean accuracy of 66% due to its ability to tolerate a reasonable degree of error in user answers. We also include the results from a number of different methods capable of estimating class conditional $p(S|c)$ to clearly illustrate the power of our random naive Bayes solution. The performances of an SVM baseline solution in addition to the discriminative random forest method in [160, 161] are also included for better comparison in table 4.7.

TABLE 4.7: Mean Accuracy of Classification Algorithms on CUB-200 Dataset

Information Source	Classification Technique	Mean Accuracy
Visual Based	SVM	19%
	Random Forest	20.51%
Textual Based	SVM	61.92%
	Random Forest	66.43%
	Naive Bayes [130]	66%
	Random Naive Bayes	68.89%

4.4.3.3 Combination Results

The fusion of information sources at input level with feature concatenation and output level with assigned equal weights show no significant difference. However, it is clear from table 4.8 that our intelligent source selection methods outperform the conventional fusion techniques. Our neural network method of intelligent source selection based on predictions from a random forest visual classifier and a random naive Bayes textual classification technique yields the best performance at 68.89%. This is in contrast to baseline results from authors in [130], who report an average accuracy of 66% based on a SVM visual classifier and a naive Bayes textual classification method.

It is also worth to mention that user responses drive up the accuracy of computer vision algorithms. Not only vision improves overall performance but also there are some cases that cannot be correctly classified without computer vision, even after asking all possible questions. The main advantage of the visual question paradigm is that contextual sources of information can easily be incorporated in the system. For instance in this dataset, information such as behaviour and habitat can be utilised as additional questions to

help with better identification of different species. It is clear that our fusion techniques improve the overall accuracy of this dataset more effectively than conventional methods.

TABLE 4.8: Mean Accuracy of Fusion Algorithms on CUB-200 Dataset

Fusion	Classifier	Concatenation		
Input Level	SVM	63.32%		
	RF	66.32%		
Fusion	Classifier	Equal Weight	Greedy Alg.	Neural Net.
Output Level	SVM+SVM	63.38%	64%	64.18%
	RF+RF	66.38%	67.3%	68.7%
	RF+RNB	66.4%	68.83%	68.89%

4.4.4 Ground Photograph Habitat Dataset

The extended version of previously deployed Ground Photograph Habitat database [179] consists of 1086 ground images with 4203 annotated polygons. There are 27 distinct habitats present in the dataset. Full details of this dataset can be explored in the preceding chapter. Evaluation metrics for multilabel classification are inherently different from those used in multiclass classification, due to the inherent differences of the classification problem. In our tests as in the previous chapter, we use the following metrics for the habitat dataset:

Hamming Loss (relaxed metric): the percentage of the wrong labels to the total number of labels. This is a loss function, so the optimal value is zero. $1 - loss$ equals to the accuracy.

Exact Match (strict metric): is the most strict metric, indicating the percentage of samples that have all their labels classified correctly.

4.4.4.1 Visual Results

We build 8976-dimensional visual feature vectors to represent the visual information of the habitat dataset. The visual features used are: Coloured Pattern Appearance Model (CPAM) [13], Geometric Blur (GB) [55], Global Image Descriptor (GIST) [171], Pyramid Histogram of Oriented Gradients (PHOG) and its variations [55], Scale-invariant Feature Transform (SIFT) and its variations, Pyramid Histogram of Visual Words (PHOW) and its variations [167], and Self-similarity Feature (SSIM) [55].

It is important to remember that the low-level visual features in this dataset particularly struggle to distinguish between semi-improved, and unimproved grassland classes of

this dataset. These classes are even subjective for human surveyors. Additionally, broad-leaved trees can be part of both the Broad-leaved Woodland habitat, which is composed of broad-leaved trees, and the Mixed Woodland habitat, which is itself composed of broad-leaved trees and coniferous trees. This similarity in classes explains the reason why the low-level features may struggle to classify these habitats. It is evident that our proposed random forest classifier outperforms the alternative baseline SVM significantly.

4.4.4.2 Textual Results

The answer to questions in table 3.17 can be used to build user response pairs suitable for the Bayesian framework of [130]. Each user response s contains an answer, and a confidence value that deals with user's uncertainty in answering the questions. The answers to the questions can also be used to build textual feature vectors of 17 dimension suitable for our approach. The results of these modelling methods can be found in table 4.9. It is again clear that our random naive Bayes method surpasses other possible solutions.

TABLE 4.9: Mean Accuracy of Classification Algorithms on Ground Photograph Habitat Dataset. Representing both Relaxed and (Strict) Metrics

Information Source	Classification Technique	Mean Accuracy
Visual Based	SVM	38.91% (3.03%)
	Random Forest	56.6% (16.78%)
Textual Based	SVM	44.22% (5.94%)
	Random Forest	52.81% (11.65%)
	Naive Bayes [130]	45.66% (6.06%)
	Random Naive Bayes	58.72% (17.94%)

4.4.4.3 Combination Results

Our proposed intelligent source selection methods prove to be very effective on this dataset. The neural network approach based on predictions from random forest visual and random naive Bayes textual classifiers achieves an intriguing mean accuracy of 68.25% and 35.19% for relaxed and strict metrics respectively. These results highlight the fact that there is an impressive 15.47% improvement over results from the same classifiers joined in an equal weight framework for the relaxed metric. This is also true for the strict metric, where a 23.25% enhancement is evident. The baseline results based on the proposed algorithm in [130] yield an average accuracy of 51.21% for relaxed and 11.65% for strict metrics.

Table 4.10 summarises the accuracies of different fusion methods. The results of both evaluation metrics we described previously is presented. As it is clear, the combination of low-level visual features with high-level knowledge of users increases the average accuracy of the algorithms. This is true with both metrics. The addition of human high-level knowledge drives up the mean accuracy of the algorithm. It is obvious that our intelligent fusion techniques are outperforming other frameworks in every aspect of the evaluation.

TABLE 4.10: Mean Accuracy of Fusion Algorithms on Ground Photograph Habitat Dataset. Representing both Relaxed and (Strict) Metrics

Fusion	Classifier	Concatenation		
Input Level	SVM	50.32% (10.78%)		
	RF	57.22% (17.94%)		
Fusion	Classifier	Equal Weight	Greedy Alg.	Neural Net.
Output Level	SVM+SVM	51.14% (11.2%)	53.3% (12.01%)	59.88% (21.63%)
	RF+RF	57.22% (17.94%)	59.7% (21.47%)	65.22% (30.78%)
	RF+RNB	52.78% (11.94%)	61.6% (22.37%)	68.25% (35.19%)

4.5 Conclusion

In this chapter, we introduced novel intelligent methodologies for selecting the most effective source of information available in human in the loop fusion frameworks. It is very interesting to note that our proposed neural network approach always produces superior, or at minimum comparable results to the greedy method of selecting information sources. It is also important to reiterate the fact that both our intelligent information fusion techniques improve classification accuracies of currently common literature methods such as: feature concatenation at input level, and considering equal weights for separate classifiers at output level [130].

We believe that our intelligent method of source selection plays a deciding role in effective incorporation of human in the loop knowledge that is truly necessary in solving difficult tasks of object recognition. Our proposed approaches effectively select the most reliable source of information from available classifiers and fuse them to produce more reliable predictions. Moreover, our introduced random naive Bayes solution to modelling user answers is a novel and efficient method in the relevant human in the loop literature. The experimental results illustrate the effectiveness of our solutions on a variety of application domains.

Chapter 5

Ranking Order of Human Information

In the previous chapters, we reviewed how to efficiently harness user abstract information for the purpose of enhancing the efficacy of classification algorithms. We believe that it is imperative to emphasise the fact that although human interactions with our proposed frameworks provide invaluable information which refines recognition outputs, the burden on the user should be kept to a minimum.

It is universally agreed that computers are very efficient machines in performing mundane tasks in contrast to humans who are good at more abstract delegations. In our proposed human in the loop applications, we utilise human knowledge in the form of answers to perceptual questions that describe complex subjects in images. Therefore, there exist settings like our selected medical applications, where a large bank of questions and answers are necessary in order to solve the problem at levels appropriate for clinical practice. It becomes obvious that in such scenarios, it is neither acceptable nor plausible to expect any user to answer hundreds of potential questions.

An intelligent sorting mechanism is vital for filtering irrelevant questions and constructing a set of decisive enquiries that improves the quality of predictions from classification algorithms. We believe that it is also expedient for any human in the loop classification algorithm to have the capability of answering those filtered questions automatically. An ideal system should be able to quantify its automatic answers with a degree of confidence, and limit user intervention to uncertain cases.

We intend to propose an innovative approach that ranks perceptual questions according to their relevance to a given image. This means that rather than asking those questions in a random stochastic order, we want to rank them based on the quality of predicted

answers and prompt merely the questions to which user responses can provide the most significant improvement in terms of accuracy. In simple words, our introduced system reduces the user's burden by minimising the number of questions necessary to achieve the best possible performance.

In this chapter, we will first review a solution to the problem of filtering relevant questions to ask from the human in the loop, and then present our intelligent random forest based technique that automatically responds to selected questions, and only invokes human engagement when it is critical. The proposed approach is potentially compatible with relevant online learning algorithms.

5.1 Problem Formulation

The principal objective is to discover appropriate methods for selecting questions related to a particular image content in order to avoid troubling the human operator constantly. We envisage that the answer to this complication is found by looking into information gain theory and several statistical methods of calculating user responses.

Formally let $Q = \{q_1, \dots, q_n\}$ be a set of possible questions, and A_i be the set of possible answers to q_i . The user's answer is therefore defined as some random variable $a_i \in A_i$. As it was described formerly, we also allow users to quantify each response with a certainty value $v_i \in V$, where $V = \{Guessing, Probably, Definitely\}$. The user's response is then a pair of random variables $s_i = (a_i, v_i)$.

We then exploit an information gain criterion and Kullback-Leibler divergence [73] to efficiently select the next set of suitable questions for a user in the loop. The upcoming question is singled out by looking into image information, its set of possible answers, as well as examining previous user responses. The question that yields maximum information gain is selected. These are the questions, which approximately divide the search span into halves in each round of algorithm iteration. The expected information gain of posing the additional question is therefore defined as the following:

$$I(c; s_i | x, S^{t-1}) = \mathbb{E}_s [D_{KL}(p(c|x, s_i \cup S^{t-1}) || p(c|x, S^{t-1}))] \quad (5.1)$$

where c is class, x is image information, and $S^{t-1} = \{s_{i(1)}, \dots, s_{i(t-1)}\}$ is the set of responses obtained by time step $t - 1$. This can be further simplified by:

$$I(c; s_i | x, S^{t-1}) = \sum_{s_i \in A_i \times V} p(s_i | x, S^{t-1}) (H(c|x, s_i \cup S^{t-1}) - H(c|x, S^{t-1})) \quad (5.2)$$

where $H(c|x, S^{t-1})$ is the entropy of $p(c|x, S^{t-1})$:

$$H(c|x, S^{t-1}) = - \sum_{c=1}^C p(c|x, S^{t-1}) \log(p(c|x, S^{t-1})) \quad (5.3)$$

These equations are employed to iteratively select a predefined number n of questions in order of importance for the user to answer. The overall logical flow of this process is illustrated in figure 5.1.

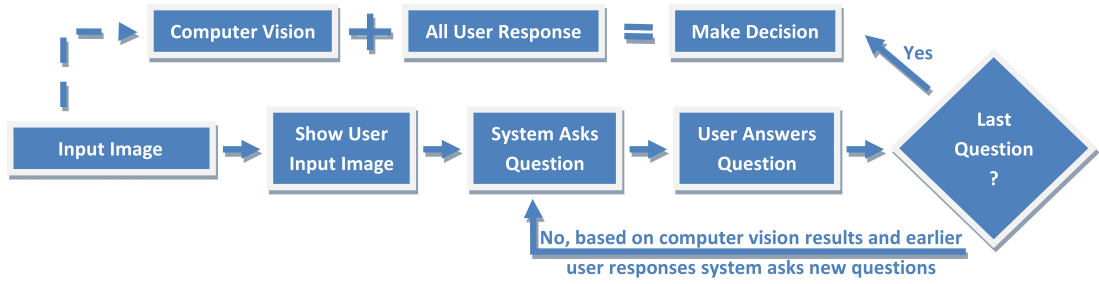


FIGURE 5.1: A sample image is displayed to the user in the loop. For each image, a question and its possible answers are also displayed. The user answers the question. The user will repeat the process until all the questions are finished or a threshold is reached. The next question is selected by looking at previous user's answers and computer vision input. The final decision is also made by combining these two elements.

The inevitable problem induced by this approach is the issue of selecting a suitable value for n , otherwise the process could in theory continue exhaustively until all possible questions in set Q are asked. The simplest solution to such complication is to learn a threshold value θ for the estimated information gain $I(c; s_i|x, S^{t-1})$ to reject statistically insignificant questions:

$$\begin{cases} I(c; s_i|x, S^{t-1}) > \theta & \text{select the next question} \\ otherwise & \text{stop asking new questions} \end{cases} \quad (5.4)$$

However, learning such a threshold is not a trivial task. By implementing this approach, the estimation of conditional probability distribution $p(c|x, S)$ is clearly dependent on the number of questions answered by the user and not the entire set of possible questions. In the next section, we will introduce our novel solution that automatically predicts answers to all available questions, and in turn enables $p(c|x, S)$ to be estimated based on a full range of possible questions and answers in any given dataset.

5.2 Automatic Answer Prediction

The performance boost by the human in the loop is only valuable if the burden on the user is kept to a minimum. The burden in our setting is defined as the number of answered questions required to reach an acceptable performance. Hence, the user is supposed to answer only a subset of questions from the entire set. Although the entire set of questions is not answered exhaustively by the user in the loop, the classification algorithm should still produce a reasonable response. The list of questions is ranked in order of importance by the information gain criterion (5.1) described previously.

Ideally, we need to automatically predict responses for those questions left unanswered by the user in a given application. This approach allows the classification algorithm to fully exploit all available perceptual questions in the set. We innovatively treat this as an image annotation problem, where predicting presence of tags is the same as guessing answers to questions. It is clear that not all automatic annotations will be perfect, and hence the least confidently predicted tags will be singled out to be asked directly from a user.

In other words, sorting the predicted probability of tags in reverse order provides the algorithm with a ranking list of most important questions to ask. We then ask the user to provide correct answers to those least confidently generated replies by the tag prediction algorithm. Thus, the user will be in charge of stopping the process at any time they deem appropriate. This in turn will reduce the number of questions to reply and therefore lowers the burden on the user involved in the decision making loop.

Our proposed approach in contrast to previous methods [130, 157] is to treat the ranking of questions in order of importance as an annotation problem, where estimating the presence of tags is the same as automatically predicting answers to perceptual questions. This approach actually enables users to answer questions as many as they desire.

5.2.1 Construction of a Random Forest for Tag Prediction

To solve this annotation problem, we propose the construction of a random forest algorithm that exploits tag information instead of the usual class labels to guide the generation of its random trees. This approach allows the correlations among different tags to be modelled implicitly. We compute the corresponding tag histograms of the left and the right child nodes after a split at each node. The tag histogram of the left node needs to be quite different from the tag histogram of the right node in order to obtain a reasonable split.

The widely adopted information gain criterion [73] is used as the score function to select a suitable split at each node in the training stage. The constructed forest's splits are therefore selected based on the tags distribution of images instead of class labels used in conventional random forests:

$$Score(split) = \Delta E = -\frac{|G_l|}{|G_n|}E(G_l) - \frac{|G_r|}{|G_n|}E(G_r) \quad (5.5)$$

where $E(G)$ is the Shannon entropy of tag distributions in the set of samples G . G_l and G_r represent the training images contained in node n 's left and right child nodes respectively. G_n is the set of training sample in node n . This random forest algorithm is trained on a collection of visual features similar to those utilised in the previous chapters. The maximum depth and the minimum number of leaf node observations are selected experimentally.

Whilst the input to the forest is a query image, which traverses down all the trained trees until leaf nodes are reached, the output from the forest is a concatenated set of training samples stored in the leaf nodes associated with the query image. The concatenated set can be represented by a histogram that highlights training samples stored in the leaf nodes, in addition to their frequency. We further clarify the usefulness of this histogram representation as the output of our constructed forest in the following section by introducing two auxiliary terms.

5.2.2 Tag Prediction by the Random Forest Algorithm

We consider the training images stored in the leaf nodes of the random forest as the “Semantic Neighbours” of the test image. From the semantic neighbours of all trees, we can conclude that the more often two images fall into the same leaf node, the more semantically similar they are, and consequently they are more likely to share similar tags. Thence, the two proposed concepts “Semantic Nearest Neighbour (SNN)” and “Semantic Similarity Measure (SSM)” literally indicate “which” and “how many times” training images fall on the same leaf node with the query image. The concepts of semantic neighbours are illustrated in figure 5.2.

More specifically, the semantic similarity measure between the test and a given training image is calculated as the number of times that the test image appears in the semantic neighbour set. Using this measure, we can sort all images in the semantic neighbours to retrieve the K semantic nearest neighbours. The K semantic nearest neighbour has an important role in prediction of tags for the test image. These predicted tags will be associated with a probability indicating how likely they are about to occur.

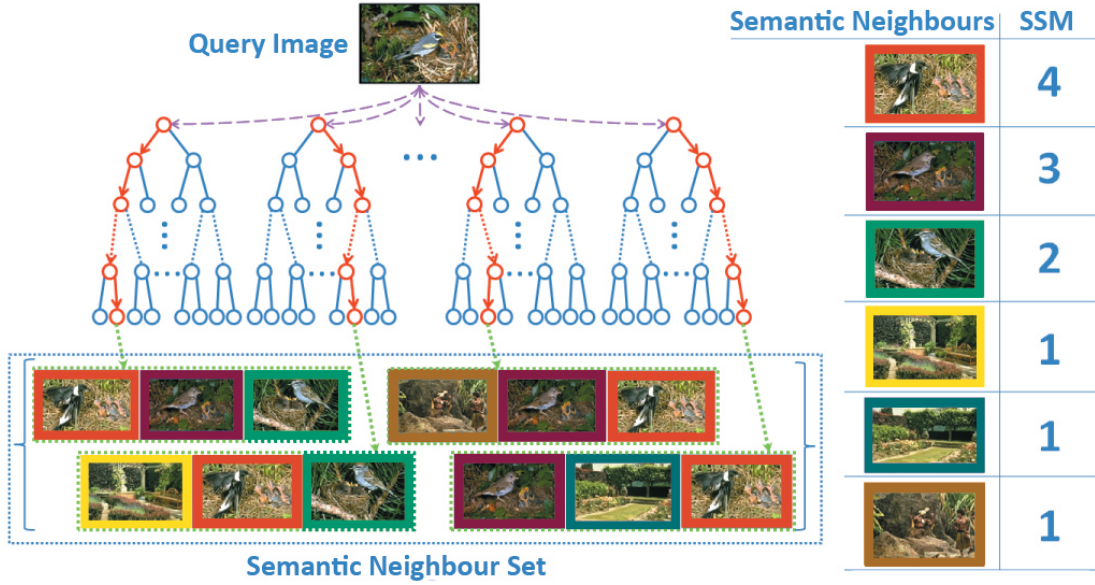


FIGURE 5.2: An example showing the concepts of semantic neighbours. A query image passes through all random trees. The training images stored at the leaf nodes on which the query image falls into form the semantic neighbour set. Based on this, the semantic similarity measure between the query and a given training image is calculated as the number of times that the given image appears in the semantic neighbour set. A larger semantic similarity measure indicates higher similarity.

Formally, we denote I the query image and Q the probabilities of assigning tags. Let I_i represents I 's i^{th} semantic neighbour with its count value denoted as c_i . The ground truth tags of I_i is denoted as T_i . Suppose there are M tags in total, hence Q and T_i can be represented as M size vectors: $Q = (q_1, \dots, q_M)^T$ and $T_i = (t_{i1}, \dots, t_{iM})^T$. Here, t_{ij} is an indicator function that shows tag j 's probability for the i^{th} image. The prediction of Q is totally influenced by T_i and c_i values:

$$q_j = \sum_{i=1}^K \left(\frac{t_{ij}}{Z} * f(c_i) \right), j \in \{1, 2, \dots, M\} \quad (5.6)$$

where Z is a normalising constant equal to $\sum_{i=1}^K \sum_{j=1}^M t_{ij}$. The term $f(c_i)$ represents a function that monotonically increases with c_i . This term in fact reflects that a neighbour with a larger count value should contribute more to the predication of tags. Based on the computed vector $(q_1, q_2, \dots, q_M)^T$, we can predict l tags for the test image which correspond to the l largest q_j values.

Possible forms of $f(c_i)$ include $f(c_i) = c_i$, $f(c_i) = c_i^2$, etc. The form of $f(c_i)$ utilised in our work is: $f(c_i) = c_i^2$, since it proves to perform empirically superior to its potential variations. However, adhoc choices of $f(c_i)$ may not be fully convincing. Authors of [126] introduce a systematic method to learn $f(c_i)$ from training data.

5.3 Experiments and Results

We evaluate our proposed algorithms on several medical datasets that contain photographic images of skin lesions from various skin conditions. The first set of experiments are carried out on the two preliminary skin datasets, which we explored previously in chapter 3. The second set of experiments are performed on our previously introduced larger “Derm2309” skin lesion dataset published in [168]. We also examine “MIAS mammographic” and “Caltech-UCSD Birds 200” datasets and present their respective results.

5.3.1 Derm90 and Derm706 Skin Conditions Datasets

In these pilot experiments, we concluded that different users could answer the same question differently but this would not affect the final correct recognition of the skin condition in a test image. This is achieved regardless of the method deployed for the estimation of conditional probability distribution $p(c|x, S)$. This classification algorithm independence is due to the influential ranking of decisive questions evoked by equation (5.1). It is obvious that the prediction of skin condition was carried out after a few time steps and not once all the questions were answered by the users.

5.3.1.1 Order of User Answers




Table 5.1 lists the orders of questions selected by the ranking equation (5.1) from “Derm90” dataset to ask from 3 different users for the purpose of classifying 3 sample test images. It is evident that albeit questions assigned to different users are selected in various orders based on their history of previous answers, and although their replies to the same questions can differ from each other, the classification algorithm is still capable of successfully classifying all test images due to presence of visual information and sensible answers. Table 5.2 lists the orders of selected questions from “Derm706” dataset in a similar experimental settings.

The correct diagnosis of Test Image 1, Test Image 2, and Test Image 3 in table 5.1 are Infantile Acne, Discoid Eczema, and Scabies respectively. It is clear from the results that our ranking algorithm (5.1) selects “Lesion Type” as the first question to ask from the users in the loop. Contextual questions such as “Site” of the affected area, and “Age” of the patient are amongst the next top queries. From the ranking algorithm’s point of view on these testing images, “Contagiousness”, and “Itchiness” are not very discriminative and therefore are not helpful in decreasing the uncertainty in labelling the samples.

Atopic Eczema, Superficial Spreading Melanoma, and Mycosis Fungoides are the ground truth labels of Test Image 1, Test Image 2, and Test Image 3 in table 5.2. It is apparent from the table that “Lesion Type” and “Surface Type” are amongst the first questions selected by the ranking algorithm discussed previously. Depending on the visual information of test images and users’ history of answers to the preceding queries, a number of questions, such as “Site”, “Arrangement”, and “Colour” of lesions, are ranked by the algorithm in order of importance in predicting the correct labels. It is clear that “Erythema” and “Duration” of the conditions are not considered to be very informative by the ranking algorithm for labelling these test samples.

Computer aided diagnosis has proved to be effective in removing the subjectivity of human observers and reducing inter-observer discrepancies in a number of medical applications [183, 184]. We believe that our results from this experiment are consistent with this conclusion.

TABLE 5.1: Order of Derm90 Questions Asked by the Ranking Algorithm from 3 different Users for 3 Test Images. The numbers refer to the questions in Table 3.3.




User 1	User 2	User 3	User 1	User 2	User 3	User 1	User 2	User 3
4	4	4	4	4	4	4	4	4
6	1	6	6	1	6	6	1	6
1	6	1	1	3	1	3	6	3
3	5	3	3	6	3	5	3	5
2	3	2	2	2	2	1	5	1
8	2	8	5	5	5	2	2	2
7	8	7	8	8	8	8	8	8
5	7	5	7	7	7	7	7	7
								
Test Image 1			Test Image 2			Test Image 3		

5.3.1.2 Frequency of User Answers

Figure 5.3 illustrates the frequency of possible answers selected by users in our pilot experiments for the “Derm706” dataset. This frequency histogram can be potentially useful in redefining the set of perceptual questions available for a skin lesion dataset.

Hypothetically, an answer that is present in almost all images of different classes cannot be very informative to a classification algorithm. This is also a valid assumption for the ranking equation (5.1), due to the underlying nature of its entropy calculation.

TABLE 5.2: Order of Derm706 Questions Asked by the Ranking Algorithm from 3 different Users for 3 Test Images. The numbers refer to the questions in Table 3.4.

User 1	User 2	User 3	User 1	User 2	User 3	User 1	User 2	User 3
9	9	9	10	10	10	9	9	9
4	3	10	9	3	4	10	1	3
10	1	6	5	1	6	6	10	10
6	6	5	6	6	3	3	5	5
3	4	3	3	5	1	1	6	1
5	10	4	1	9	5	12	3	6
1	5	1	13	11	9	5	12	13
13	7	13	4	12	11	13	4	11
11	2	11	8	4	12	11	13	4
7	13	12	11	2	2	8	11	12
2	11	2	12	13	13	4	7	2
12	12	8	2	8	7	7	8	8
8	8	7	7	7	8	2	2	7
								
Test Image 1			Test Image 2			Test Image 3		

Informally this can be explained by the fact that the less likely an event is, the more information it provides when it occurs.

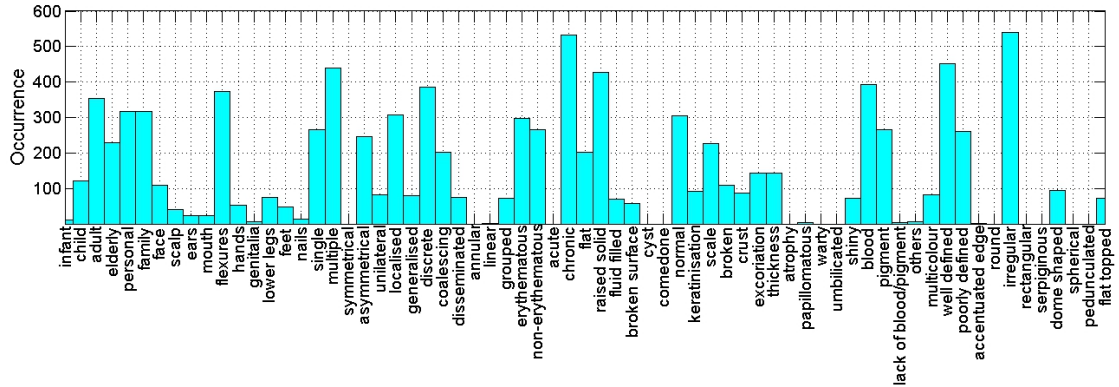


FIGURE 5.3: Frequency of answers selected by users in the “Derm706” dataset.

5.3.2 Derm2309 Skin Conditions Dataset

We evaluate the efficacy of our proposed random forest approach for automatic prediction of answers to perceptual questions on our larger “Derm2309” skin conditions dataset. There are several intriguing conclusions that can be drawn from this experiment:

5.3.2.1 Automatic Answers Accuracy

It is very interesting to note that our solution is capable of answering all the questions automatically, and achieving a superior performance to results based solely on visual descriptors. In chapter 3, we revealed the advantages of incorporating high-level user information into the conventional vision algorithms. It became clear that textual descriptors built based on user provided information could improve the average accuracy of our selected classification algorithms. It is evident in this experiment that automatic answers to the same perceptual questions can replace real users' responses to some degree and still improve the overall performance of the classification algorithm.

Classification accuracy based on visual features saturates at 15.76% on this dataset, whilst the combination of these visual features with our fully predicted answers, where no real user is involved in answering the questions, results in an average accuracy of 17.91%. These results are obtained from a standard random forest classification algorithm similar to our introduced technique in the previous chapters.

5.3.2.2 Questions Ranking Effect

It is imperative to clarify the fact that users in our proposed system do not need to answer all questions. Our model utilises both user provided answers, as well as automatically predicted tags in calculating the final classification results, despite the fact that some of these answers may have been wrongly predicted. Figure 5.4 represents the effect of adding user provided answers to our solution. As we gradually replace least confident automatic tags with user provided answers, the average accuracy rises. It is important to note that the system does not require to utilise all user tags to achieve its peak performance. In the same figure, results from randomly picked tags are also presented for comparison purposes. It is obvious that randomly picking user tags has not the same effective results as selecting the least probably correct ones using our solution.

Our proposed method reaches the peak performance on this dataset after utilising 30 answers from the user. However, it is clear that the same results cannot be achieved by replacing the automatically predicted answers with real user responses in a random stochastic order. This reiterates the fact that our proposed method of reducing user's burden can be fairly effective.

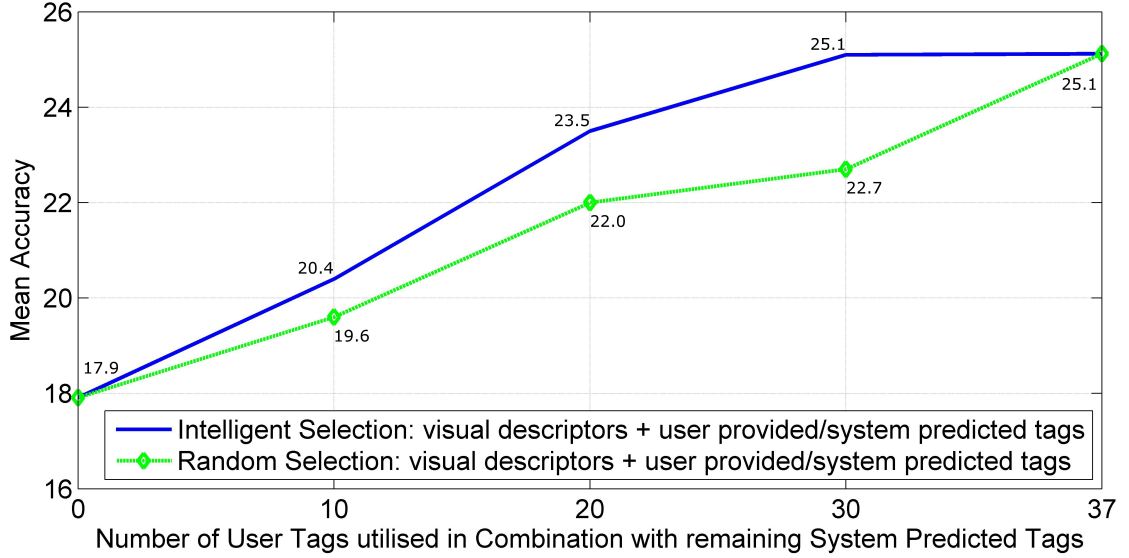


FIGURE 5.4: Mean classification accuracy results on Derm2309 Dataset: System predicted tags reduce the number of user tags required to achieve peak performance. Results from randomly picked tags is also illustrated.

5.3.3 MIAS Mammographic Dataset

As a further evaluation of our proposed random forest approach for automatic prediction of answers to perceptual questions, we set up a similar experiment to our previous test on the MIAS Mammographic Dataset.

5.3.3.1 Automatic Answers Accuracy

Classification accuracy based purely on visual descriptors levels at 28.44% on this dataset. This result is achieved using a random forest algorithm as the method of classification. The combination of visual features and our automatically predicted answers, where no real user is involved in answering the questions, results in a mean accuracy of 44.39%.

This highlights the fact that even a fully automatic set of answers can significantly increase accuracy rates of a very difficult visual object recognition application. In chapter 3, we extensively discussed the benefits of high-level user information. It is once again observed that textual descriptors built based on user provided information can improve the average accuracy of selected classification algorithms.

5.3.3.2 Questions Ranking Effect

Figure 5.5 illustrates the effect of introducing user provided answers to our random forest classification algorithm. As before, we gradually replace least confident automatic tags with user provided answers. It becomes evident that the average overall accuracy increases. It is imperative to emphasize once more that the system does not need to deploy all user tags to achieve its peak performance. In the same figure, results from randomly picked tags are also presented. It is again obvious that randomly picking user tags has not the same effect as selecting the least probably correct ones using our proposed solution. Our model utilises both user provided answers, and automatically predicted tags in calculating the final classification results. This is despite the fact that some of these tags may have been inaccurately predicted.

Our proposed ranking algorithm reaches the peak performance on this dataset after utilising 8 answers from the user. It is evident that similar results cannot be obtained by replacing the automatically predicted answers with real user responses in a random order. This reiterates the fact that our proposed method of reducing user’s burden can be useful in interactive medical applications that may require a large bank of questions and answers.

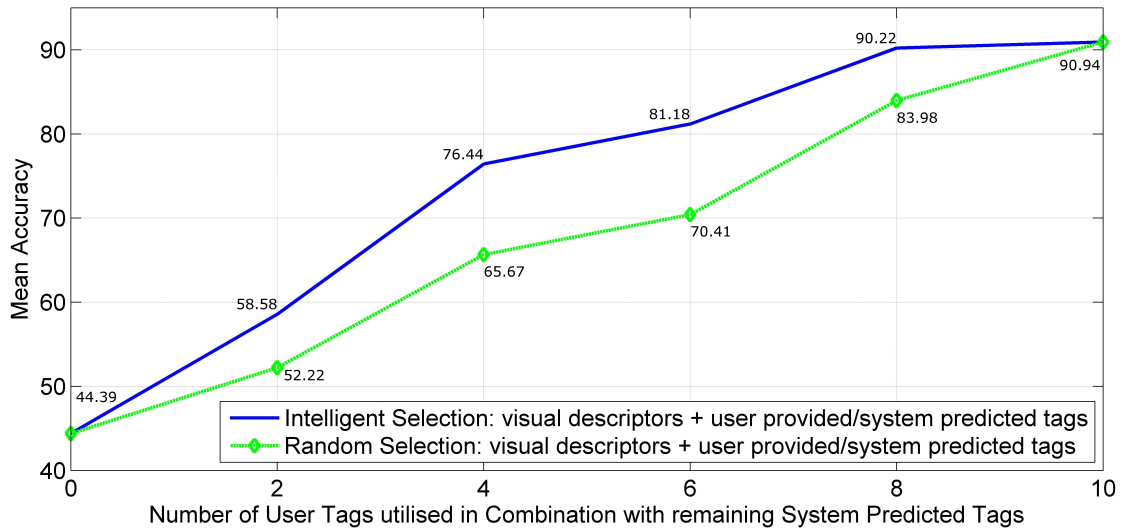


FIGURE 5.5: Mean classification accuracy results on MIAS Dataset: System predicted tags reduce the number of user tags required to achieve peak performance. Results from randomly picked tags is also presented.

5.3.4 Caltech-UCSD Birds 200 Dataset

We attempted to examine our automatic prediction technique on the “Caltech-UCSD Birds 200” Dataset. However, the achieved results were not as promising as our tests

on medical datasets. We assume that this failure is partially due to the low number of training images in the dataset, considering its large variability of class labels and possible tags. We believe that a more effective method of automatic annotation is needed to obtain reasonable results on this dataset.

5.3.4.1 Automatic Answers Accuracy

Classification accuracy based on visual descriptors levels at 20.51% on this dataset, whilst the combination of the same visual descriptors with our automatically predicted answers, where no user is involved in answering the questions, leads to an accuracy of 21.94%. These results are obtained from a standard random forest classification algorithm similar to our introduced technique in the previous chapters.

Although the overall outcome of our experiments on this dataset fails to show the encouraging results we observed on the preceding datasets, it is still worth noting that our proposed solution is capable of answering all the questions automatically, and achieving a slightly better performance than results based only on visual features. As it was mentioned before, in chapter 3, we revealed the advantages of incorporating user information into vision algorithms. It is again observable that textual descriptors improve the average accuracy of our classification algorithm. Automatic answers to the perceptual questions can replace real users' input to some extent and marginally improve the overall performance on this dataset.

5.3.4.2 Questions Ranking Effect

Our model utilises both user provided answers and automatically predicted tags in calculating the final classification results. Figure 5.6 represents the effect of adding user provided answers to our solution. As we gradually replace least confidently predicted tags with user provided answers, the average accuracy rises. It is desirable for us to design a system that reaches its peak performance without utilising all user answers. In the same figure, results from randomly selected tags are also presented.

It is clear that our proposed method of replacing system predicted tags fails to achieve an improved result over a random method of replacement on this dataset. The random method achieves its peak performance at approximately 65% after utilising 230 user provided tags, whereas our intelligent method levels at 61.13% after incorporating the same number of user tags. It takes the entire set of user answers for our proposed method to reach the peak performance. Therefore, this failure prevents the desired alleviation of burden imposed on our users in the loop.

In contrast to the last two cases, the total number of tags to predict is larger in this dataset. We believe that our intelligent selection method fails to accomplish its purpose due to the smaller number of training samples per class, and the larger number of class labels present in this dataset. More training samples and a more robust method of label prediction should in theory help with fixing the issues faced on this dataset.

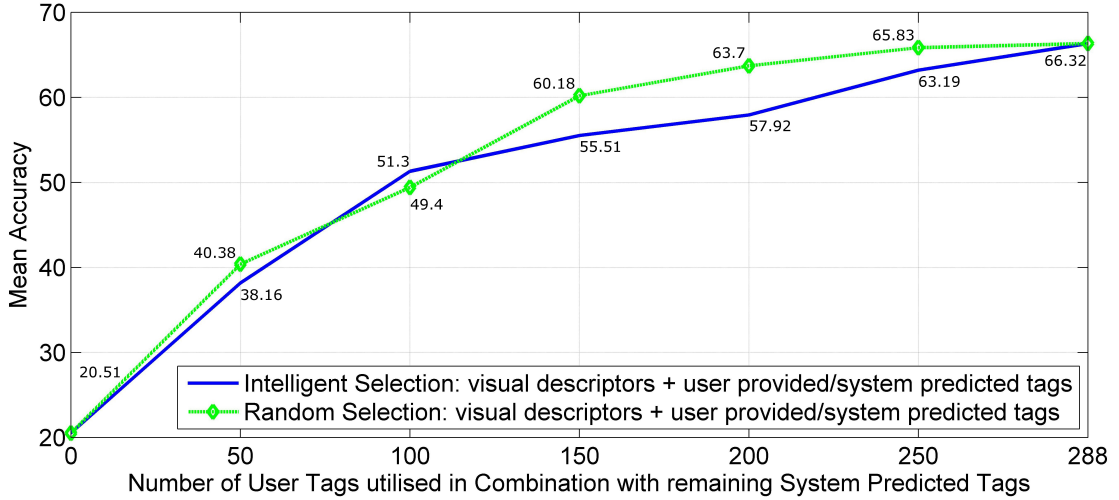


FIGURE 5.6: Mean classification accuracy results on CUB-200 Dataset: System predicted tags should reduce the number of user tags required to achieve peak performance. Results from randomly picked tags is illustrated.

5.4 Conclusion

In this chapter, we introduced intelligent methods to select the best sequence of questions that improves classification performance, and simultaneously removes the burden from user's side. Firstly, we introduced the information gain criterion useful for effective ranking of questions in order of their decisiveness in finalising a correct classification. Secondly, we proposed our random forest approach for automatic prediction of user answers in order to eradicate the need for exhaustive examination of every single question in a given dataset. We tested the efficacy of these algorithms in a medical setting on several skin conditions datasets with various levels of size and difficulty.

In the following chapter, we will discuss an innovative approach that separates the most discriminative features and has the capability to enhance the accuracy of many classification algorithms, including our interactive techniques. The process of selecting a subset of relevant features is vital to designing robust human in the loop vision models. Our proposed selection techniques in the next chapter eliminate redundant or irrelevant visual and textual features, and therefore they are considered to be an important primary step in utilising any of the introduced techniques in the previous chapters.

Chapter 6

Discriminative Subspace Selection based on Mutual Information

In many machine vision and pattern recognition applications, the infamous “curse of dimensionality” is a well-known problem. A widely used approach to alleviate this complication is subspace methods, where the original data is projected onto a new space in which lower dimensional feature vectors are used to approximate the original vectors. Amongst conventional subspace methods are: Principal Component Analysis (PCA) [185], Linear Discriminant Analysis (LDA) [186], various frequency analysis based transforms such as the Fourier Transform (FT) [187], and its derivative the Discrete Cosine Transform (DCT) [188], short-time Fourier Transform, Wavelet Transform (WT) [189], and other variations of frequency analysis method such as the Hadamard Transform (HT) [190]. Random Projection (RP) [191], where the original data is projected onto a lower dimensional random directions, has also been used for dimensionality reduction.

Nevertheless, it is reasonable to claim that all these subspace methods were not originally developed specifically for pattern recognition or object classification applications. PCA is for identifying subspaces, in which the input data has the largest variance, such that the inverse transform from a lower dimensional subspace recovers the original data with minimum loss of energy. FT, DCT, WT, and HT are all for retaining the lower frequency components of the original data. Surprisingly, pattern recognition literature conventionally adopts these methods, as they were originally developed for dimensionality reduction, without questioning if they also make theoretical and practical sense when applied to pattern recognition.

We believe an ideal representation is in a space where the classes of data are well separable. As we will demonstrate later, directly applying these methods in pattern

recognition or classification is not always the best practice, and a new information theory based method for selecting the subspaces can enhance the performance of a learning system substantially. A discriminative subspace can also boost the efficiency of “Human in the Loop” classification algorithms, which may suffer from the abundance of potential visual and textual descriptors. This chapter makes the following main contributions:

1. It enhances a common practice widely used by practitioners in the field of pattern recognition. To the best of our knowledge, this work originally highlights the interesting fact that in implementation of dimensionality reduction subspace methods, such as [192–196] for pattern recognition or classification applications, practitioners should not directly adopt the conventional methods but instead explicitly opt for a discriminative subspace from the transforms.
2. It develops an information theory based technique for systematically selecting subspaces that are discriminative and therefore are suitable for pattern recognition or classification purposes.
3. It presents extensive experimental results on a variety of computer vision and pattern recognition tasks to illustrate that the subspaces selected based on the maximum mutual information criterion will almost always improve performance regardless of the classification techniques in use.

In the rest of this chapter, our setup is the regular multiclass setting, where we have a labelled dataset $\{(x_i, y_i) \in X \times Y\}$ sampled *iid* from a distribution D on $\mathbb{R}^d \times [l]$. We therefore need a classifier $f : \mathbb{R}^d \rightarrow [l]$ with low generalisation error $\mathbb{P}_D(f(x) \neq y)$. To keep focus on the effectiveness of our subspace selection method, we restrict ourselves to three classifiers: Random Forest, Support Vector Machine, and Naive Bayes in the evaluation section of this chapter. We believe our discriminative dimensionality reduction method can also improve the performance of other classifiers given any particular settings.

6.1 Problem Formulation

Our proposed technique is close in nature to existing methods that work by finding suitable subspaces constructed from data. These methods generally find directions v that maximise a signal to noise ratio:

$$R(v) = \frac{v^T S v}{v^T N v} \quad (6.1)$$

where matrices S and N are selected such that quadratic forms $v^T S v$ and $v^T N v$ represent signal and noise respectively along direction v . This ratio allows us to categorise similar methods in those that can produce many directions, and those that can generate discriminative directions. One of the most straightforward statistics involving both features and labels to extract directions is the matrix $\mathbb{E}[xy^T]$. This is the collection of class-conditional mean feature vectors in a multiclass classification setting. However, it is relatively safe to expect that such simple first moment statistics fail to contain all the information available in the data distribution. Alternatively, a collection of the conditional second moment matrices $C_i = \mathbb{E}[xx^T | y = i]$ can be used to extract features. Nevertheless, there is no reason to expect that these extracted directions are specific to class i . The directions may be very similar for all classes, and hence not very discriminative. A simple solution to this problem is to use the ratio of expected projection magnitudes conditional on different class labels. This necessitates to address which class pairs to extract directions. When the number of class labels is modest, it is possible to consider all ordered pairs of classes but unfortunately this is not the case in many applications.

We believe that it may be advantageous to explore higher than second order statistical information to derive a discriminative subspace, which not only enables low dimensional representation of inputs but also allows input projections to be well-separated. For instance, kernel based subspace methods [33, 34, 197] exploit higher order statistics to derive a subspace. Information theory [198] can also be used to benefit from higher order statistics. Mutual information measures general statistical dependence between variables rather than their linear correlations. It is also invariant to monotonic transformations performed on the variables. These illustrate a number of advantages that information theoretic approaches may have over similar methods for deriving discriminative subspaces.

6.2 Mutual Information Subspace

Inspired by the aforementioned advantages of mutual information over alternative solutions, we introduce the implementation of a method based on information theory. This technique exploits mutual information to guard against selecting non-discriminative directions, while allowing the extraction of a diverse range of transformation vectors. Formally, we let X and Y be discrete random variables with sets of possible outcomes. We then define the mutual information between X and Y as:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (6.2)$$

where $p(x, y)$ is the joint probability distribution function of X and Y , and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Mutual information measures the amount of information x conveys about y . In the context of object classification, Fano's inequality [35] provides us with a lower bound for the probability of error, and an upper bound for the probability of correct classification. Formally, the probability of misclassification error $P_e = P(y \neq y')$ has the following bound:

$$P(y \neq y') \geq \frac{H(Y) - I(X; Y) - 1}{\log(C)} \quad (6.3)$$

where $H(Y)$ is the entropy of Y , X is the ensemble of random variable x , and C is the number of object classes.

Equation (6.3) quantifies at best how well we can classify objects using available features. However, an upper bound of the probability of misclassification error cannot be expressed in terms of Shannon's entropy. The best one can do is to minimise the lower bound to ensure an appropriately designed classification algorithm performs well. Since both C and $H(Y)$ are constants, we can maximise the mutual information $I(X; Y)$ to minimise the lower bound of the probability of misclassification error. At this point, the task develops into finding the transform function that minimises this lower bound. We therefore implement the preceding transform function by finding a low-dimensional representation f of the original set of N d -dimensional observations $X_{d \times N}$. This is achieved by projecting the original input data onto a k -dimensional ($k \ll d$) subspace using a $k \times d$ matrix G :

$$f_{k \times N} = G_{k \times d} X_{d \times N} \quad (6.4)$$

In this chapter, we are motivated by mutual information and an information theoretic criterion to select the projection matrix G^* :

$$G^* = \arg \max_{\forall G} I(GX; Y) \quad (6.5)$$

where Y is the identity variable of input variable X , $I(X; Y)$ is the mutual information between X and Y .

The mutual information $I(GX; Y)$ is calculated by estimating the probability density from a finite number of samples. Let us assume that we have N number of samples in the training set. The probability densities $p(x)$, $p(y)$, and $p(x, y)$ can be approximated by histograms. The difference between the true value \bar{I} and the estimation I of the mutual information can be estimated by adapting the analysis of [199], as the following:

$$\Delta I \equiv I - \bar{I} \approx \frac{1}{2N} \left(\sum_{x,y} \frac{(\delta n_{xy})^2}{n_{xy}} - \sum_x \frac{(\delta n_x)^2}{n_x} - \sum_y \frac{(\delta n_y)^2}{n_y} \right) \quad (6.6)$$

where the sums are over the discretised intervals and δn are the fluctuations of the countings with respect to the mean values ($\delta n = n - \bar{n}$). The approximation is valid up to the second order of the relative fluctuations, and if the ratios do not change significantly with x and y .

Different subspace methods differ in their way of computing and selecting base vectors of the projection matrix G . We need to clarify that our criterion for selecting base vectors in the projection matrix G differs intrinsically from their conventional counterparts. Specifically, we want to employ the maximum mutual information criterion (6.5) to select the appropriate k base vectors.

To find the first base vector of G , we select one computed vector from a subspace method at a time, and project all other computed vectors from the training set onto that selected vector. The projections are a set of scalar numbers, which can be discretised. The samples' identities can be used to estimate the joint probability. The joint probability can be deployed to estimate the mutual information between the projections and the class distribution, as discussed previously. The vector with projection outputs that maximises the mutual information is selected as the first transform base of matrix G . This base is subsequently removed from the vectors' set.

The process continues until all required k base vectors are found. If we have a large enough pool of samples, it is reasonable to assume that most informative representative bases will be selected. The representational quality and discriminative power of f is dependent on the computed base vectors of matrix G . In the rest of this chapter to clarify the practicality of this approach, we exemplify the computation and selection procedures of a data-independent, data-dependent, and the random projection methods of dimensionality reduction using mutual information criterion. The pseudo code in 3 summarises our process described in this section thus far.

Algorithm 3 Mutual Information Subspace Algorithm**Require:** Observations: $X_{d \times N}$, Labels: Y , Number of base vectors: k Step 0: Compute subspace transformation matrix G e.g. eigenvectors of covariance matrix form $G_{d \times d}$ in PCAStep 1: Compute projections of samples X onto base vectors of G i.e. form projection matrix $Z_{d \times N} = GX$ Step 2: Compute mutual information for every base vector of Z as in eq.6.2i.e. calculate $I(Z_{i \times 1}; Y), \forall i \in d$ Step 3: Sort all base vectors based on their calculated $I(Z_{i \times 1}; Y)$ i.e. construct matrix $G^* = \emptyset$ **while** there exist unsorted base vectors v **do** $v = \arg \max_{i \in d} I(Z_{i \times 1}; Y)$ $G^* = G^* \cup \{v\}, Z = Z - \{v\}$ **end while****return** First k rows of G^* **6.2.1 Examples of Common Subspace Methods****Data Independent Transform - DCT:**

The projection matrix G in the Discrete Cosine Transform (DCT) method of dimensionality reduction is the transform coefficients. Conventionally, reduction is achieved in the inverse transform by discarding the transform coefficients corresponding to the highest frequencies. In contrast to the convention, we propose to use the mutual information criterion (6.5) to select the k transform coefficients used in the projection matrix, and not simply the coefficients corresponding to the lowest frequencies.

Data Dependent Transform - LDA:

LDA computes an optimal projection by minimising the within-class distance and maximising the between-class distance simultaneously, thus achieving maximum class discrimination. The optimal transformation in LDA can be readily computed by applying an eigendecomposition on the so-called scatter matrices. More specifically, eigenvectors corresponding to the $k - 1$ largest eigenvalues form columns of G . Instead of relying on largest eigenvalues to form the projection matrix, our proposed mutual information criterion selects the base vectors of G .

Data Dependent Transform - PCA:

In Principal Component Analysis (PCA), eigenvalue decomposition of data covariance matrix is computed as $\mathbb{E}\{XX^T\} = E\Lambda E^T$, where the columns of matrix E are the eigenvectors of data covariance matrix $\mathbb{E}\{XX^T\}$ and Λ is a diagonal matrix containing the respective eigenvalues. The k eigenvectors corresponding to the k largest eigenvalues

of the covariance matrix form the projection matrix G . In contrast to this traditional approach of selecting the first k vectors, we rely on the mutual information criterion (6.5) explained previously to select the appropriate bases of matrix G .

Random Projection - RP:

A simple probability distribution can form the base vectors of the projection matrix G in the Random Projection method of dimensionality reduction:

$$g_{ij} = \begin{cases} +1 & \text{with probability } 1/3 \\ 0 & \text{with probability } 1/3 \\ -1 & \text{with probability } 1/3 \end{cases} \quad (6.7)$$

Conventionally, the first k computed vectors from this distribution constitute the projection matrix G . However as before, we propose to use the mutual information criterion (6.5) to select the required k bases instead.

6.2.2 Useful Properties

The feature descriptors resulting from maximising equation (6.2) have a number of useful properties that we list below:

Proposition (Maximum Dependence) By maximising equation (6.2), we ensure that two random variables X and Y are statistically as dependent as possible. This means that feature vectors most relevant to a certain class is always preferred.

Proof. Mutual information $I(X; Y) = 0$ if and only if X and Y are independent random variables. In such case, the joint probability between the two variables is $p(x, y) = p(x)p(y)$, and therefore:

$$\log \left(\frac{p(x, y)}{p(x)p(y)} \right) = \log 1 = 0 \quad (6.8)$$

This criterion enables our algorithm to maximally exploit the data by selecting the most informative descriptors of a certain class.

Proposition (Nonlinear Separation) Mutual information ability to consider nonlinear relations between variables can be advantageous over linear methods of analysis like correlation.

Proof. Mutual information is capable of measuring general dependence between two variables. Variables x and y are linearly independent if $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$, and generally independent if $p(x, y) = p(x)p(y)$. Hence, general independence implies linear independence, but not vice versa. This property enables this algorithm to be advantageous over linear methods of analysis.

We intend to test our introduced method on data-independent, data-dependent, and the random projection dimensionality reduction techniques. We prove that our solution empirically works well, as we illustrate in the evaluation section of this chapter.

6.3 Experiments and Results

We evaluate our discriminative subspace selection method on a number of benchmark datasets. However, we firstly assess simple synthetic data to graphically illustrate the efficacy of our proposed technique. All results presented in this section are based on a 5-time repeated random sub-sampling cross validation method. We fix three commonly used multiclass classifiers and compare their outcomes to solely examine the performance of our algorithm and discard other potentially influential factors. The selected classifiers are an ensemble of 200 bagged decision trees, an RBF kernel SVM, and a naive Bayes model.

It is imperative to note that the main focus here is not to exactly achieve state-of-the-art classification performance on all datasets through a vigilant engineering procedure, but to emphasise the usefulness of our mutual information technique given any method of subspace dimensionality reduction. Our algorithm outperforms the original subspace methods in all tests or at minimum produces comparable results. State-of-the-art performance is easily attainable once our feature selection approach is combined with carefully crafted descriptors and fine-tuned hyperparameters of robust classifiers.

6.3.1 Synthetic Data

In this pilot experiment, we generate a 100-by-2 matrix R of random variables chosen from a multivariate normal distribution with mean vector μ , and a symmetric positive semi-definite covariance matrix Σ to simplify the visualisation process. The synthetic data is an example of a binary classification problem, where two sets of 50 instances belong to 2 discriminable class labels. The data is randomly split into a 50:50 training and testing sets.

We project the 2- d features onto a 1- d subspace and use a naive Bayesian classifier to categorise the data in the projected 1- d space. The purpose of this experiment is to verify the soundness of our method and to demonstrate the possible advantages of our technique over conventional subspace methods. We hence compare our algorithm with the well-known subspace method of Principal Component Analysis, and measure the mutual information between the projected 1- d features and the data's class labels to examine the relation between mutual information and classification error.

Figure 6.1 depicts a situation, where projecting the data onto the 1- d PCA base fails to discriminate between the two classes, whilst projecting the data onto our subspace alleviates the classification problem by making classes easier to separate. In this example, PCA finds the direction of maximal variance but fails to determine the most discriminative direction. From this simple experiment, we also conclude that the mutual information and classification error have a direct relation: the higher the mutual information contained in the subspace, the lower the classification errors are and vice versa.

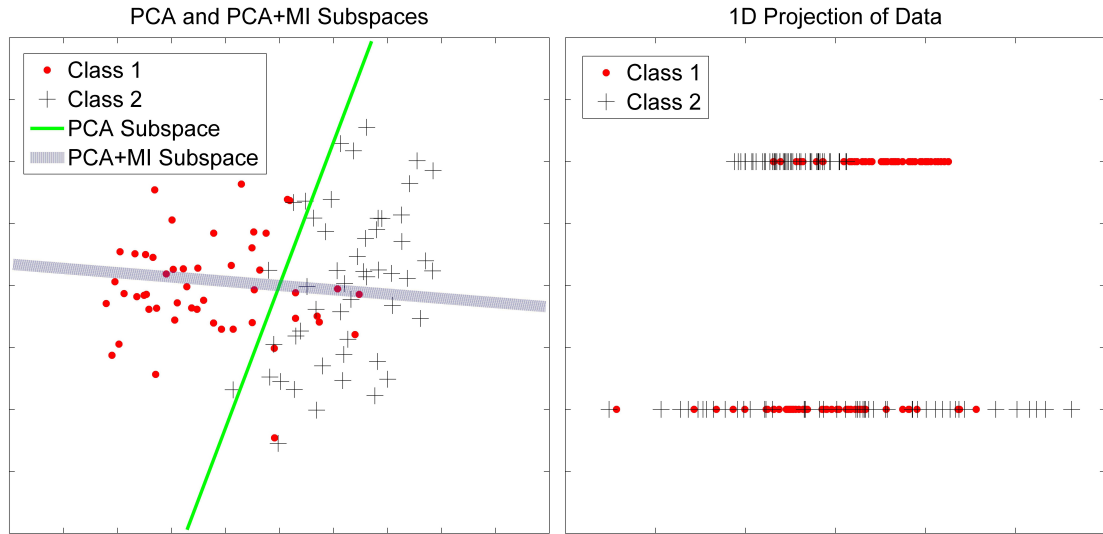


FIGURE 6.1: 2D illustration of a two class dataset, where PCA and mutual information subspaces have been drawn. The projection on a 1D line clearly demonstrates the separability of data using our subspace selection method (right box, top line).

Principal Component Analysis is a widely used linear transform for dimensionality reduction. It is an optimal reduction technique in the mean square error sense. The eigen subspace captures the directions of maximal variance in data. Nevertheless, as we just illustrated in figure 6.1, the maximal variance and discriminative directions are not guaranteed to coincide. Therefore, PCA subspace is not always appropriate for representing the data in a classification settings. We believe that our algorithm captures higher order, more general statistical information, and therefore is a more suitable candidate than the alternative solutions.

6.3.2 Derm2309 Skin Conditions Dataset

This familiar dataset [168] contains images of skin conditions from 44 different diseases. There are 880 training and 1429 testing images, totalling 2309 images in the dataset. In the original release of this dataset, there are 20 training images per class, and the rest are used for testing. The sheer difficulty of this dataset, in addition to our adherence to its original split of training and testing sets lead to the observed low average classification accuracies in our experiments.

Table 6.1 represents the mean classification accuracies based on different fractions of the data's original dimensions. The mean accuracies of the same classifiers using no dimensionality reduction technique on this dataset are 20.37%, 19.4%, and 18.88% for random forest, SVM, and naive Bayes respectively. It is evident from the LDA results that classification of this dataset can be performed in the reduced space more accurately than in the original space. This improvement is observable from the outputs of all the three classifiers.

The LDA method achieves the best performance on this dataset by using 70% of the available data. This result is obtained by the random forest classifier. It is clear that our mutual information technique can indeed enhance the accuracies returned by all the dimensionality reduction methods. Our technique enhances the accuracies of LDA and random projection more noticeably than the other dimensionality reduction methods using the random forest and the naive Bayes classifiers.

These results are based on PHOW-HSV, and textual features enclosed in the public release of this dataset.

TABLE 6.1: Mean Accuracies in Percentage on Derm2309 Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	16.52	16.03	16.79	15.89	14.75	14.75	15.75	15.75	12.27	12.97	11.99	11.64
DCT+MI	16.66	17.49	16.79	17.91	16.75	16.75	17.75	17.75	14.38	14.45	13.38	13.45
LDA	12.06	15.12	20.90	25.40	14.84	15.32	19.26	20.23	10.96	12.67	16.74	20.41
LDA+MI	15.21	19.74	23.97	26.87	17.43	19.01	19.75	20.23	14.25	17.67	19.32	20.41
PCA	13.37	20.64	18.54	16.72	17.06	19.65	18.25	17.27	12.32	15.52	16.98	17.00
PCA+MI	15.26	21.20	19.17	18.61	19.30	19.86	18.25	17.34	13.04	17.01	17.84	17.07
RND	11.24	12.08	12.43	13.06	16.17	15.75	15.75	15.75	10.22	12.25	14.83	15.81
RND+MI	14.46	14.60	14.46	15.44	16.17	15.82	15.75	15.75	13.43	14.76	15.81	15.95
W/O Reduction	20.37				19.40				18.88			

6.3.3 MIAS Mammographic Dataset

We test our proposed subspace selection algorithms on the aforementioned MIAS database released by the Mammographic Image Analysis Society [174]. Full details of our experiment's setup for this dataset was described in the previous chapters. It may be necessary to mention that adopting a smaller number of trees in this chapter leads to the observed lower average classification accuracies in our experiments.

Table 6.2 represents the mean classification accuracies based on different fractions of data's original dimensions. The average classification accuracies using no dimensionality reduction technique on this dataset are 35.41%, 33.22%, and 32.14% for the random forest, SVM, and naive Bayes classifiers respectively. The results of all our proposed dimensionality reduction methods illustrate that classification of this dataset can be performed in the reduced space more accurately than in the original space. This enhancement is observable from the outputs of all the three classifiers.

The PCA and random projection methods enhanced by our mutual information technique achieve the best performance on this dataset. These enhancements are observable by all the three classifiers. In spite of our mutual information technique's ability to improve the average accuracies of all the dimensionality reduction methods, LDA illustrates the smallest enhancement in the results obtained by all the classifiers. The random naive Bayes classifier demonstrates a modest improvement of results by using any dimension larger than 30% of the data's original dimensionality. This is mostly noticeable with DCT and LDA techniques.

These results are based on visual feature vectors of length 5756, which include: SIFT [11], grey PHOG features [167], Grey-Level Co-occurrence Matrix (GLCM) [175], Local Binary Patterns (LBP) [167], Local Phase Quantisation (LPQ) [176], and Canny Edge Detector [177] features. To utilise human in the loop information, the dataset questions were exploited to build textual descriptors.

6.3.4 MSRC 21-class Dataset

MSRC 21-class is a well-known dataset [5] that contains 591 images. Each image has pixel-level ground-truth labels from 21 semantic classes. These 591 images are split into 276 for training, 59 for validation, and the remaining 256 images for testing purposes.

Table 6.3 represents the mean classification accuracies based on different fractions as before. The mean accuracies of the same classifiers using no dimensionality reduction technique on this dataset level at 64.4%, 61.95%, and 59.3% for random forest,

TABLE 6.2: Mean Accuracies in Percentage on MIAS Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	29.74	30.17	31.47	31.47	26.29	27.43	29.22	32.10	26.72	29.31	30.60	30.60
DCT+MI	35.78	36.64	37.50	37.07	28.90	31.19	35.29	36.61	29.74	30.60	32.33	31.47
LDA	30.43	30.86	38.71	39.74	30.95	30.05	36.17	37.78	23.71	25.17	30.07	30.16
LDA+MI	34.60	34.54	39.64	40.33	33.06	33.19	38.10	38.66	26.12	28.43	31.81	32.92
PCA	32.33	34.48	31.47	26.29	31.44	33.63	32.11	30.03	31.47	31.03	31.90	30.17
PCA+MI	38.79	35.78	43.10	39.22	36.72	35.09	40.79	36.71	32.76	33.19	33.62	34.48
RND	30.60	29.74	24.14	33.62	26.64	27.01	26.02	29.82	30.17	30.17	31.03	31.03
RND+MI	41.38	42.67	39.66	43.97	36.54	36.78	38.48	39.97	36.64	35.34	34.05	34.91
W/O Reduction	35.41				33.22				32.14			

SVM, and naive Bayes respectively. It is observable from the results of all our employed dimensionality reduction techniques that classification of this dataset cannot be performed in the reduced space more accurately than in the original space. This is noticeable from the outputs of all the three classifiers. Dimensionality reduction is primarily used for compression. Thus, it can only help learn a better classifier, when the data does have a low dimensional structure. We believe that the observed difference between the classification results in the original and the reduced space requires further investigation in the future, as the gap between the two seems to be abnormal on this dataset.

The DCT and random projection methods achieve the best performance in the reduced space on this dataset by using 50% and 70% of the data's original dimensions. These results are obtained by the random forest and the SVM classifiers. LDA's results do not exhibit a substantial improvement after utilising 70% of the available data. This is due to the absence of a significant difference between the ranking of eigenvalues and eigenvectors returned by the original LDA and our mutual information technique. PCA also demonstrates a similar results to some extent on this dataset using 70% of the data's original dimensions.

These results are based on Texton, colour histograms, and PHOG visual features.

6.3.5 Oxford Flower Recognition Dataset

The Oxford flowers dataset [200] contains 17 different types of flowers. Each class contains 80 samples, 40 for training, 20 for validation, and the rest for testing.

Table 6.4 describes the mean classification accuracies based on different fractions of data's original dimensions. The mean classification accuracies using no dimensionality

TABLE 6.3: Mean Accuracies in Percentage on MSRC 21-class Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	51.94	56.13	57.74	57.10	52.26	52.26	56.26	56.26	40.32	39.68	41.94	44.19
DCT+MI	54.84	56.13	58.06	59.35	54.26	54.26	59.26	59.26	42.58	46.45	47.42	47.10
LDA	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	41.29	46.77	46.77	46.77
LDA+MI	51.94	54.19	52.26	50.00	51.94	54.84	52.90	50.00	41.29	50.32	50.32	46.77
PCA	56.77	53.55	46.13	46.77	50.65	49.03	49.68	50.00	42.58	47.74	46.13	40.97
PCA+MI	60.32	57.74	54.52	47.42	52.26	49.35	49.68	50.00	43.23	48.06	46.45	41.94
RND	50.00	50.97	52.90	54.19	52.26	52.26	52.26	52.26	37.10	42.58	42.90	43.55
RND+MI	54.19	58.06	56.45	59.68	58.26	58.26	58.26	58.26	40.65	42.90	45.48	45.81
W/O Reduction	64.40				61.95				59.30			

As some results seem unintuitive even though confirmed through repeated experiments, we acknowledge that further analysis of the results is warranted as part of our future work for this chapter.

reduction technique on this dataset are 49.53%, 46.4%, and 43.35% for the random forest, SVM, and naive Bayes classifiers respectively. It is evident from the LDA results that classification of this dataset can be performed in the reduced space more accurately than in the original space. This improvement is observable from the outputs of the random forest and the SVM classifiers. PCA results produced by the SVM and the naive Bayes classifiers also demonstrate an improvement in the reduced space in comparison to the original space.

The LDA method enhanced by our mutual information technique achieves the best performance on this dataset by using 50% or 70% of the available data. These results are obtained by the SVM classifier. The naive Bayes classifier illustrates a steady improvement of about 1% in the results of all the dimensionality reduction techniques. The noticeable exception to these improvements is the DCT method, which benefits from a larger enhancement using all fractions of the data's original dimensions. Our proposed technique enhances the accuracies of random projection more noticeably than the other dimensionality reduction methods using the random forest classifier. LDA and SVM contrastingly exhibit some of the smallest improvements on this dataset.

The visual features used are: HSV colour histograms, SIFT, and MR8 texture descriptors.

6.3.6 Pascal VOC2007 Challenge Dataset

Pascal visual object classes of 2007 challenge dataset [201] has 20 distinguishable classes, as follows: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car,

TABLE 6.4: Mean Accuracies in Percentage on Oxford Flower Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	41.76	43.53	42.65	42.94	41.76	39.41	39.71	39.41	31.18	29.71	29.12	28.24
DCT+MI	45.00	43.53	42.94	46.18	41.76	39.41	39.71	39.41	33.12	34.29	35.76	36.06
LDA	37.94	45.88	47.65	49.71	40.59	47.94	48.53	50.00	31.76	38.24	38.82	40.00
LDA+MI	39.12	46.47	48.82	50.00	40.59	47.94	50.00	50.00	31.76	38.24	40.00	40.00
PCA	44.71	40.59	35.29	32.94	46.76	46.76	45.88	45.88	47.65	42.35	36.18	39.71
PCA+MI	47.94	42.06	41.18	35.00	47.06	48.76	47.88	47.88	48.24	43.53	36.18	39.71
RND	38.82	40.88	39.41	42.06	44.12	40.88	39.12	40.88	35.88	39.41	40.59	42.35
RND+MI	41.18	43.53	47.65	42.94	44.41	40.88	39.71	41.18	36.76	40.00	41.47	42.65
W/O Reduction	49.53				46.40				43.35			

motorbike, train, bottle, chair, dining table, potted plant, sofa, and tv/monitor. Train, validation, and test sets have 9963 images in total containing 24640 annotated objects.

Table 6.5 displays the average classification accuracies based on different fractions of original data. The average classification accuracies using no dimensionality reduction technique on this dataset stand at 31.41%, 29.82%, and 27.28% for the random forest, SVM, and naive Bayes classifiers respectively. It is evident from the DCT, LDA, and PCA results that classification of this dataset can be performed in the reduced space more accurately than in the original space. This improvement is observable from the outputs of all the three classifiers. The only exception on this dataset is the random projection technique that illustrates a less accurate results in the reduced space. This outcome is evident across all the three classifiers.

The DCT and PCA methods enhanced by our mutual information technique achieve the best performance on this dataset. These results are obtained by the random forest and the SVM classifiers. A steady improvement of approximately 2% is observable from all the dimensionality reduction methods using the same classifiers. The naive Bayes classifier demonstrates the least significant enhancement in all the dimensionality reduction techniques, particularly using any fractions larger than 50% of the data's original dimensions.

The results are based on 15 publicly released visual descriptors: Gist, DenseSift, DenseSiftV3H1, HarrisSift, HarrisSiftV3H1, DenseHue, DenseHueV3H1, HarrisHue, HarrisHueV3H1, Rgb, RgbV3H1, Lab, LabV3H1, Hsv, and HsvV3H1.

TABLE 6.5: Mean Accuracies in Percentage on Pascal VOC2007 Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	32.45	32.33	32.18	32.96	30.50	30.96	32.49	30.96	25.37	26.31	26.24	26.62
DCT+MI	34.24	33.55	33.71	33.49	33.20	31.55	33.49	33.18	27.72	27.03	27.71	27.90
LDA	25.61	26.96	28.02	30.08	26.55	28.31	30.90	32.80	23.43	25.55	28.78	29.24
LDA+MI	27.33	27.96	30.50	33.99	29.72	30.83	31.55	33.55	25.50	28.18	28.78	29.24
PCA	32.77	32.74	30.96	30.83	30.78	31.46	30.02	31.96	29.53	28.94	28.06	27.99
PCA+MI	34.78	34.49	33.46	32.08	32.80	34.49	33.96	33.49	29.72	29.22	28.25	28.28
RP	21.42	21.92	28.61	22.02	20.77	22.33	28.18	21.24	20.94	21.12	21.12	20.84
RP+MI	22.80	23.43	29.43	25.30	21.18	24.55	29.71	24.08	22.66	22.22	21.66	21.66
W/O Reduction	31.41				29.82				27.28			

6.3.7 UCI Machine Learning Repository Datasets

We further present experiments on two UCI repository datasets available from: [202]. Sonar, Mines vs. Rocks is the dataset used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network [203]. The dataset's task is to discriminate between sonar signals bounced off a metal cylinder and those bounced off a cylindrical rock.

Table 6.6 describes the mean classification accuracies of UCI-Sonar dataset based on different fractions of original dimensions. The average classification accuracies using no dimensionality reduction technique on this dataset are 76.87%, 77.3%, and 74.64% for the random forest, SVM, and naive Bayes classifiers respectively. It is evident from the results of all our selected dimensionality reduction techniques that classification of this dataset can be performed in the reduced space more accurately than in the original space. This improvement is observable from the outputs of the random forest and the SVM classifiers.

The LDA method enhanced by our mutual information technique obtains the best performance on this dataset. These results are achieved by the random forest and the naive Bayes classifiers. The most significant enhancement by our proposed technique can be observed in the results of the random projection method. Smaller improvements are evident from the results of other dimensionality reduction methods using fractions larger than 30% of data's original dimensions.

There is only one feature used for the purpose of classification in this dataset. Each feature is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time.

TABLE 6.6: Mean Accuracies in Percentage on UCI-Sonar Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	79.55	82.95	82.95	81.82	77.73	77.73	77.73	77.73	61.36	57.95	56.82	56.82
DCT+MI	81.82	82.95	82.95	88.64	81.73	81.73	81.73	81.73	62.50	60.23	60.23	56.82
LDA	87.50	87.50	87.50	87.50	87.50	87.50	87.50	87.50	83.77	83.77	85.77	85.77
LDA+MI	89.77	87.50	87.50	87.50	87.50	87.50	87.50	87.50	87.77	87.77	89.77	89.77
PCA	76.14	77.27	75.00	75.00	79.55	81.82	77.27	79.55	73.86	68.18	67.05	69.32
PCA+MI	82.95	82.95	78.41	76.14	80.68	82.95	79.55	80.68	76.14	71.59	68.18	70.45
RND	54.55	72.73	73.86	77.27	47.73	67.73	73.86	73.86	54.55	53.41	59.09	60.23
RND+MI	76.14	80.68	80.68	81.82	57.95	67.73	77.73	77.43	75.00	75.00	73.86	76.14
W/O Reduction	76.87				77.30				74.64			

Multiple Features dataset consists of features of handwritten numerals ('0'-'9') extracted from a collection of Dutch utility maps. 200 patterns per class, for a total of 2000 patterns, have been digitised in binary images. These digits are represented in terms of the following six feature sets: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages in 2 by 3 windows, 47 Zernike moments, and 6 morphological features.

Table 6.7 also illustrates the mean classification accuracies of UCI-MFeat dataset based on different fractions of data's original dimensions. The mean accuracies of the same classifiers using no dimensionality reduction technique on this dataset are 96.51%, 94.51%, and 93.38% for random forest, SVM, and naive Bayes respectively. It is clear from the results of all our selected dimensionality reduction techniques that classification of this dataset can be performed in the reduced space more accurately than in the original space. This improvement is evident from the outputs of the random forest and the SVM classifiers.

The average enhancement of the DCT technique using our mutual information method is about 1%. This is observable from the results of all the three classifiers. LDA's improvements are only significant in results obtained from a 10% fraction of the data. PCA and random projection methods both exhibit small improvements on most fractions of the data's original dimensions. The random projection method enhanced by our mutual information technique achieves the best performance on this dataset by using 70% of the available data, and by utilising the random forest classifier.

These classification results are based on the 6 publicly released features, discussed previously.

TABLE 6.7: Mean Accuracies in Percentage on UCI-MFeat Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	96.25	96.50	97.00	97.00	90.75	90.75	93.75	93.75	86.75	88.50	89.25	89.25
DCT+MI	96.75	96.75	97.00	97.00	91.75	91.75	95.75	95.75	90.50	90.75	91.00	91.75
LDA	61.25	95.00	97.00	98.00	64.50	95.75	98.00	98.00	65.25	96.25	97.00	98.00
LDA+MI	65.75	95.25	97.00	98.00	71.75	96.25	98.00	98.00	76.50	97.50	98.00	98.00
PCA	97.25	97.00	96.75	95.00	79.75	88.25	94.50	95.00	96.75	95.50	95.00	94.00
PCA+MI	97.50	97.50	97.00	97.50	79.75	91.25	94.50	98.00	96.75	96.00	95.25	94.00
RND	94.00	96.00	97.00	96.50	52.75	80.75	90.75	90.75	93.75	95.00	95.50	95.00
RND+MI	96.50	98.00	97.50	98.50	54.50	80.75	96.75	96.75	93.75	95.75	95.50	96.25
W/O Reduction	96.51				94.51				93.38			

6.3.8 Yale Face Recognition Dataset

The Yale Face database [196] contains 165 grayscale images of 15 individuals in GIF format. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

Table 6.8 illustrates the mean classification accuracies based on different fractions of data's original dimensions. The average accuracies of the same classifiers using no dimensionality reduction technique on this dataset level at 85.85%, 84.92%, and 83.65% for random forest, SVM, and naive Bayes respectively. It is observable from the results of our employed dimensionality reduction techniques that classification of this dataset cannot be always performed in the reduced space more accurately than in the original space. It is only our introduced PCA technique that improves the average performances of the random forest and the naive Bayes classifiers.

It seems that our proposed mutual information technique does not significantly influence the final accuracy results of the dimensionality reduction methods. This is evident when any fractions larger than 30% of the data's original dimensions are utilised by the three classifiers. The most noticeable improvement is observed from the DCT method using a 10% fraction. This is due to the fact that DCT selects the lowest frequencies for construction of the transformation matrix. It is apparent that the selected frequencies are not necessarily discriminative for the purpose of classification on this dataset.

The only visual feature used in this dataset is the image intensity values ranging from 0 to 255.

TABLE 6.8: Mean Accuracies in Percentage on Yale Face Dataset

Classifier	Random Forest				SVM				Naive Bayes			
Dimension	10%	30%	50%	70%	10%	30%	50%	70%	10%	30%	50%	70%
DCT	64.44	66.67	68.89	68.89	67.78	65.56	65.56	65.56	37.78	51.11	51.11	46.67
DCT+MI	68.89	68.89	68.89	68.89	76.67	67.78	65.56	65.56	71.11	60.00	51.11	46.67
LDA	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00
LDA+MI	82.22	82.22	82.22	80.00	82.22	82.22	82.22	80.00	82.22	82.22	82.22	80.00
PCA	82.22	77.78	68.89	53.33	75.56	73.33	71.33	71.33	62.22	82.22	71.11	64.44
PCA+MI	88.89	82.22	80.00	68.89	79.56	77.33	71.33	71.33	66.67	86.67	75.56	71.11
RND	75.56	80.00	75.56	75.56	71.11	74.44	70.00	70.00	71.11	73.33	73.33	73.33
RND+MI	77.78	82.22	77.78	77.78	71.11	76.67	70.00	70.00	73.33	73.33	77.78	75.56
W/O Reduction	85.85				84.92				83.65			

6.3.9 Interpretation of Results

We believe our proposed method can improve the performance of any subspace procedure. The computational expense of DCT is $O(dN \log_2(dN))$. LDA's $O(d2N)$ calculation is dominated by the computation of the within-class scatter and its inverse. PCA is estimated as $O(d^2N) + O(d^3)$, and Random Projection complexity is $O(dkN)$. We know that the computational overhead from our algorithm on these subspace methods is negligible. It is not feasible to interpret a certain message from the results of random projection technique on our datasets, apart from a solid improvement over the original method. However, a few interesting points can be observed from the results on the DCT, LDA, and PCA procedures:

First Few Low Frequency Bases are not Necessarily Discriminative

The results of our DCT experiments on the selected datasets illustrate the point that the first few low frequency bases are not necessarily preferred by our mutual information method. Interestingly, 31.44% of the first 10 frequencies selected by our mutual information algorithm are different from the original DCT technique.

First Few Principal Components are Discriminative

Unlike DCT, the results from all datasets in our experiments highlight the fact that the first 10 principal components are also usually selected by our mutual information method. There is only a 3.38% difference between the selection of our mutual information method, and the original PCA technique. The same conclusion cannot be drawn from the largest eigenvalues and their corresponding eigenvectors returned by the LDA approach. There appears to be a significant difference between typical LDA selections and our proposed mutual information method.

Performance of Algorithms Exhibits Asymptotic Behaviour at 70%

As the number of base vectors in the projection matrices increases, i.e. the dimension of input data escalates, the difference in classification accuracies of our mutual information method and the original subspace techniques becomes negligible. This behaviour starts to surface for DCT, LDA and PCA procedures with approximately 70% of original dimensions. Table 6.9 displays these asymptotic results by listing the number of different bases returned by the two methods in percentage.

TABLE 6.9: Difference in Bases returned by our MI and the Conventional Methods in Percentage

	Dim.	Derm	MIAS	MSRC	OXF	VOC	Sonar	MFeat	Yale	Avg.
DCT	10%	78.30	88.89	87.72	55.45	37.16	66.67	48.44	83.33	68.24
	30%	63.21	59.26	69.19	52.42	32.55	61.11	43.81	60.53	55.26
	50%	44.15	42.22	54.36	41.09	26.75	43.33	40.43	34.38	40.83
	70%	26.01	25.60	31.67	26.62	17.96	26.19	26.65	22.47	25.39
LDA	10%	20.00	40.11	10.00	06.39	13.16	06.00	100.0	14.29	26.24
	30%	38.46	38.97	33.33	05.09	11.75	12.40	33.33	14.29	23.45
	50%	09.09	16.78	28.57	12.50	12.34	12.40	02.33	14.29	13.53
	70%	03.09	13.10	00.00	03.00	10.02	04.01	00.00	00.00	04.15
PCA	10%	12.83	15.22	16.14	12.75	12.97	20.00	04.50	11.96	13.29
	30%	11.89	10.83	11.45	11.52	13.02	17.65	08.19	08.56	11.63
	50%	06.57	21.70	13.52	10.73	10.34	03.45	04.85	09.33	10.06
	70%	05.54	22.43	09.77	10.39	04.71	04.88	08.52	10.33	09.57

6.4 Conclusion

In this chapter, we illustrated a novel method based on mutual information for discriminative subspace selection. We demonstrated empirical efficacy via multiple experiments on different datasets. Due to this empirical, computational, and statistical properties, we believe our proposed model has the potential capacity to be employed in a wide range of computer vision and pattern recognition problems including human in the loop algorithms.

Chapter 7

Concluding Remarks

In this thesis, we studied semantic image understanding with human involved in the decision making loop of vision algorithms. This chapter will summarise our major contributions, highlights their limitations, discuss potential improvements, and directions for future work. We will conclude with a summary.

7.1 Main Contributions

In chapter 3, we introduced a novel “Random Forest” based human in the loop framework that efficiently fuses visual features of images with user provided information. This approach enables fast prediction and superior classification performance on a number of human in the loop datasets. User abstract knowledge in this method is harnessed in the shape of user answers to perceptual questions. These responses are used to build “Textual Descriptors” that are compatible with random forest classifiers. Contrary to generative Bayesian frameworks in the following chapter, this is a direct discriminative approach that leads to data fusion at input level of classifiers.

Our next contribution was described in chapter 4, where we proposed a “Random Naive Bayes” model of capturing human high-level information that is compatible with generative human in the loop Bayesian frameworks. We additionally introduced innovative “Human in the Loop Fusion Schemes” that intelligently select the most effective source of information available for making predictions. Through experiments on a variety of human in the loop datasets, we demonstrated the advantages of our “Random Naive Bayes” model in comparison to the-state-of-the-art methods of capturing user abstract knowledge both in terms of accuracy and efficiency. We also stated that our new methodologies for intelligent selection of information sources

outperform their competitors in tasks such as fine-grained categorisation. This was achieved by devising two separate learning algorithms that assign variable weights to each source of information, unlike conventional methods which assume all sources are equally effective in determining the correct class label in classification settings.

Chapter 5 revealed new methods to reduce unnecessary human intervention in decision making procedures. Our proposed algorithm determines the most “Efficient Sequence of Information” to obtain from users in the decision making loop in order to minimise their unnecessary involvement in mundane tasks. Our approach in practice allows users to be more concerned with abstract functions instead. This was accomplished first by examining information theory and leveraging a criterion [130] that ranks perceptual questions in order of their importance in arriving at the correct classification rapidly. We then scrutinised algorithms that take this approach further by attempting to predict answers to the perceptual questions automatically without the need for human intervention. This became possible by treating the issue in hand as an automatic annotation problem, where the algorithm seeks human assistance merely in uncertain cases.

We demonstrated a novel remedy for the “curse of dimensionality” in pattern recognition problems that is based on “Mutual information and Fano’s Inequality” methods in chapter 6. Our approach separates the most discriminative descriptors and has the ability to enhance the accuracy of many classification algorithms. The process of selecting a subset of relevant features is critical for designing robust human in the loop vision models where the vast availability of options in selecting visual or textual descriptors make it difficult to find and adopt the most effective settings. In other words, our devised selection techniques simply eliminate redundant or irrelevant visual and textual features. The evaluation results confirmed the fact that our proposed algorithm is capable of enhancing classification accuracies regardless of decisions made in selecting the dimensionality reduction method or the classification algorithm.

Finally in appendix A, we cited our published paper that extensively describes our procedures for collecting and constructing a human in the loop adaptable dataset, which contains 2309 photographic images of 44 different skin conditions. The publicly released version of our dataset contains the extracted visual features from images, in addition to users’ answers that we harnessed using the “Amazon Mechanical Turk” interface. We believe that our skin conditions dataset is very useful in facilitating the development of computer aided medical diagnostic techniques in dermatology.

7.2 Limitations and Future Work

We truly believe that the work presented in this thesis is merely a start in their respective fields. We think that they have either raised new questions to investigate, or left room for future improvements.

In chapter 3, we introduced a simple method to harness human abstract knowledge in form of answers to predefined perceptual question. An interesting but difficult subject to explore is the matter of evaluating the possibility of constructing questions automatically based on information provided by visual descriptors. A short examination of the relevant literature reveals a preliminary work [128] based on an interactive approach that intelligently selects discriminative regions of an image to be named as meaningful or meaningless by the human operator. We speculate the selected discriminative and nameable regions of images can then be used to derive more relevant perceptual questions depending on the application's domain. A comprehensive set of evaluations on more detailed "Questions and Answers" bank should turn the possibility of developing practical solutions more plausible. Furthermore, the uncertainties in user answers to the perceptual questions were modelled using several adhoc values listed in table 3.1. A more efficient approach should learn these values, which practically influence the outcomes of classifiers, from a training set of data collected from human in the loop participants.

Bayesian frameworks in chapter 4 were implemented by making a few assumptions that turned the estimation of full joint distributions easier. Obviously, more sophisticated techniques should be exploited to estimate the probability distribution $p(c|x, S)$. In our preliminary work, this was not plausible due to insufficient training data. The performance of these algorithms and subsequently the classifiers is directly dependent on the quality of visual or textual descriptors. We believe that the current move towards deep learning methods for feature extraction rather than handmade descriptors can play an imperative role for the efficiency of our developed human in the loop algorithms. We further assume that expanding our experiments on a wider range of classification algorithms can draw a more vivid picture on the effectiveness of our introduced fusion techniques.

The question ranking technique based on information gain presented in chapter 5 is a greedy algorithm that may fail to find the global optimum. There should be alternatives that alleviate this issue. Our introduced technique of image annotation in the same chapter can also be enhanced to perform more accurately. There are still dark corners that need immediate attention in order to accomplish a robust but practical human in the loop framework. One of the main difficulties is the issue of selecting an appropriate number of questions to ask from users. There are also user behaviour analysis techniques

that could potentially lead to more effective implementation of user interfaces. These improved interfaces will help to harness operators’ abstract knowledge more efficiently. Furthermore, the negative results we observed on the “Caltech-UCSD Birds 200” dataset need detailed investigation.

Chapter 6 technique of discriminative subspace selection is a competent method of enhancing classification accuracies. The main drawback of our current solution is the issue of tractability. Although our proposed mutual information method’s complexity is negligible, it is directly affected by the subspace method in use and its computational costs are therefore proportional with sample size. As computers are getting more powerful, it is not unrealistic to assume that our algorithm becomes computationally practicable on very large datasets. Additional systematic evaluation of our mutual information technique on more variety of object classification and pattern recognition applications can further reassure the efficacy of our proposed algorithm.

Last but not least, we truly think that our developed skin conditions “Derm2309” dataset in appendix A can be improved by devising a bigger “Questions and Answers” bank. Involving a larger group of users with various prior medical knowledge would definitely elevate the usability of our dataset in facilitating the development of computer aided medical diagnostic solutions.

7.3 Epilogue

In this thesis, we aimed to provide practical solutions to the enduring problem of visual content understanding by incorporating human high-level information in the decision making loop using simple methodologies. Whilst chapter 3 enclosed the most straightforward technique of utilising both visual and textual information in a discriminative random forest framework, chapter 4 revealed a number of generative techniques that model each source of information individually in order to determine the most effective source for the purpose of class label prediction. Chapter 5 presented methodologies to reduce human unnecessary involvement in mundane tasks by only focusing on cases where their invaluable abstract knowledge is of utter importance. Finally, the feature selection algorithm presented in chapter 6 is in fact the primary step in any of the object classification algorithms. Hence, its influence on improvement of precisions for various human in the loop algorithms can prove to be integral.

We truly assume that it is intriguing to conclude this thesis with a little food for thought:

Should the community merely seek the path of finding automatic resolutions to the problem of visual content understanding, and consider

human in the loop techniques solely as an intermediate step that meanwhile enables implementation of practical solutions to many real world problems?

If autonomous solutions are the long seeking answers to many real world problems, how can we ease people's reaction to such proposals?

Or there actually exist critical applications where individuals will never trust fully autonomous solutions with no human supervision, and hence further enhancement of these techniques is a must?

Appendix A

Derm2309 Skin Conditions Dataset

Automatic recognition of skin conditions from medical images is still an unreachable goal at the current level of technology. By involving human in the loop and combining high-level cognitive information with traditional low-level visual features, developing practically useful machine vision technologies suitable for medical applications may become a reality.

In this appendix, we provide the URL to our challenging human in the loop image recognition dataset [168], which not only provides imagery data but also the images' associated high-level information.

We believe that our dataset is a very useful addition to the current computer-aided diagnosis systems within medical imaging groups, as well as more general visual object recognition communities.

A.1 Data Format

All our dataset materials are saved as Matlab “mat” files. The public release of files includes: image URLs, bounding boxes information, extracted PHOW visual features, user provided answers, and a list of skin conditions classes.

A.2 Method to Read

All data can be loaded into the Matlab workspace using the Matlab “load” command.

A.3 Evaluation Criteria

Average classification accuracy is the main evaluation criterion for our multiclass dataset. Precision and recall may also be calculated for each individual class present in the dataset.

A.4 Download Address

Dataset is available to download from: <https://db.tt/RuPsutgR>

Bibliography

- [1] P. McCorduck. *Machines who think: a personal inquiry into the history and prospects of artificial intelligence*. AK Peters Series. A.K. Peters, Natick, Massachusetts, USA, 2004.
- [2] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5:1–5:60, May 2008.
- [3] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.
- [4] G. Qiu and P.C. Yuen. Editorial: Interactive imaging and vision-ideas, algorithms and applications. *Pattern Recognition*, 43(2):431–433, February 2010.
- [5] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In A. Leonardis, H. Bischof, and A. Pinz, editors, *European Conference on Computer Vision (ECCV)*, volume 3951 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2006.
- [6] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.
- [7] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1582–1596, September 2010.
- [8] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision (IJCV)*, 62(1-2):61–81, April 2005.
- [9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, June 2005.

- [10] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conference on Computer Vision (ECCV)*, volume 6315 of *Lecture Notes in Computer Science*, pages 352–365. Springer Berlin Heidelberg, 2010.
- [11] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, November 2004.
- [12] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 244–252. Curran Associates, Inc., December 2010.
- [13] G. Qiu. Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition*, 35(8):1675–1686, August 2002.
- [14] A. Gersho and R.M. Gray. *Vector quantization and signal compression*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers, Boston, USA, 1992.
- [15] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1784, June 2011.
- [16] F. Perronnin, Yan L., J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3384–3391, June 2010.
- [17] L. Li, H. Su, E.P. Xing, and L. Fei-fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386, 2010.
- [18] S. Sadanand and J.J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1241, June 2012.
- [19] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014.
- [20] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, December 1989.

- [21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.
- [22] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Machine Learning (ICML)*, pages 81–88, New York, NY, USA, 2012.
- [23] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates, Inc., 2012.
- [24] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1915–1929, August 2013.
- [25] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1470–1477, October 2003.
- [26] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision (ECCV)*, pages 589–600, 2006.
- [27] L. Zheng, G. Qiu, J. Huang, and H. Fu. Salient covariance for near-duplicate image and video detection. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 2537–2540, September 2011.
- [28] R. Behmo, N. Paragios, and V. Prinet. Graph commute times for image representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.
- [29] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006.
- [30] A. Lucchi, Y. Li, X. Boix, K. Smith, P. Fua, and E. BIWI. Are spatial and global constraints really necessary for segmentation? In *IEEE International Conference on Computer Vision (ICCV)*, pages 9–16, 2011.
- [31] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(3): 226–239, March 1998.

- [32] H. Fu, G. Qiu, and H. He. Feature combination beyond basic arithmetics. In J. Hoey, S. McKenna, and E. Trucco, editors, *The British Machine Vision Conference (BMVC)*, pages 58.1–58.11. BMVA Press, 2011.
- [33] P. Zhang, J. Peng, and C. Domeniconi. Kernel pooled local subspaces for classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(3):489–502, June 2005.
- [34] S.W. Kim and B.J. Oommen. On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods. In T.T.D. Gedeon and L. Fung, editors, *AI 2003: Advances in Artificial Intelligence*, volume 2903 of *Lecture Notes in Computer Science*, pages 783–795. Springer Berlin Heidelberg, 2003.
- [35] R.M. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press Classics. MIT Press, Cambridge, MA, USA, 1961.
- [36] J.W. Fisher and J.C. Principe. A methodology for information theoretic feature extraction. In *IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence*, volume 3, pages 1712–1716, May 1998.
- [37] T. Butz and J.P. Thiran. Multi-modal signal processing: An information theoretical framework. Technical report, Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), 2002.
- [38] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 281–288, October 2003.
- [39] G. Qiu and J. Fang. Classification in an informative sample subspace. *Pattern Recognition*, 41(3):949–960, March 2008.
- [40] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.
- [41] D. Keren. Painter identification using local features and naive bayes. In *16th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 474–477, 2002.
- [42] M.N. Prasad, A. Sowmya, and I. Koch. Feature subset selection using ica for classifying emphysema in hret images. In *17th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 515–518, August 2004.

- [43] Y. Liu, J. Guo, and J. Lee. Halftone image classification using lms algorithm and naive bayes. *IEEE Transactions on Image Processing*, 20(10):2837–2847, October 2011.
- [44] H. Tabia, M. Gouiffes, and L. Lacassagne. Motion histogram quantification for human action recognition. In *21st International Conference on Pattern Recognition (ICPR)*, pages 2404–2407, November 2012.
- [45] E. Parzen. On estimation of probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.
- [46] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.
- [47] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.
- [48] J.C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.
- [49] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, January 2004.
- [50] G. Mehmet. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, July 2011.
- [51] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, October 2007.
- [52] F. Orabona, J. Luo, and B. Caputo. Online-batch strongly convex multi kernel learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–794, June 2010.
- [53] A. Zien and C.S. Ong. Multiclass multiple kernel learning. In *International Conference on Machine Learning (ICML)*, pages 1191–1198, New York, NY, USA, 2007.
- [54] Y. Yan, R. Rosales, G. Fung, M.W. Schmidt, G.H. Valadez, L. Bogoni, L. Moy, and J.G. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. *Journal of Machine Learning Research - Proceedings Track*, 9:932–939, May 2010.

- [55] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 606–613, 2009.
- [56] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler. lp norm multiple kernel fisher discriminant analysis for object and image categorisation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3626–3632, June 2010.
- [57] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(11):1475–1490, November 2004.
- [58] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.
- [59] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.
- [60] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.
- [61] B. Widrow and M.E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record, Part 4*, pages 96–104, New York, 1960.
- [62] F. Rosenblatt. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Report (Cornell Aeronautical Laboratory). Spartan Books, Washington, USA, 1962.
- [63] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*. MIT Press, Cambridge, MA, USA, 1986.
- [64] M.T. Hagan, H.B. Demuth, and M. Beale. *Neural Network Design*. PWS Publishing Co., Boston, MA, USA, 1996.
- [65] R. Socher, E.H. Huang, J. Pennin, C.D. Manning, and A. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 801–809, 2011.
- [66] D.C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649, 2012.

- [67] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.
- [68] R.E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. Adaptive computation and machine learning. MIT Press, Cambridge, MA, USA, 2012.
- [69] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, USA, 1984.
- [70] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [71] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, October 1997.
- [72] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.
- [73] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.
- [74] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.
- [75] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, October 2007.
- [76] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1022–1029, June 2009.
- [77] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, June 2011.
- [78] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pages 96–103, New York, NY, USA, 2008.
- [79] T. Sharp. Implementing decision trees and forests on a gpu. In *European Conference on Computer Vision (ECCV)*, volume 5305 of *Lecture Notes in Computer Science*, pages 595–608. Springer, 2008.

- [80] N. Payet and S. Todorovic. $(rf)^2$ - random forest random field. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1885–1893. Curran Associates, Inc., 2010.
- [81] G. Fanelli, J. Gall, and L. van Gool. Real time head pose estimation with random regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 617–624, June 2011.
- [82] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1668–1675, November 2011.
- [83] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 415–422, October 2011.
- [84] A. Prinzie and D. van den Poel. Random multiclass classification: Generalizing random forests to random mnl and random nb. In R. Wagner, N. Revell, and G. Pernul, editors, *Database and Expert Systems Applications*, volume 4653 of *Lecture Notes in Computer Science*, pages 349–358. Springer Berlin Heidelberg, 2007.
- [85] M. Godec, C. Leistner, A. Saffari, and H. Bischof. On-line random naive bayes for tracking. In *20th International Conference on Pattern Recognition (ICPR)*, pages 3545–3548, August 2010.
- [86] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1792–1799, November 2011.
- [87] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.
- [88] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1062–1069, June 2010.
- [89] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 32(9):1627–45, September 2010.

- [90] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 28(4):594–611, April 2006.
- [91] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 380–387, 2005.
- [92] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.
- [93] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 237–244, September 2009.
- [94] Y.J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2010.
- [95] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2579–2586, 2011.
- [96] P.J. Schuffler, T.J. Fuchs, C.S. Ong, V. Roth, and J.M. Buhmann. Computational tma analysis and cell nucleus classification of renal cell carcinoma. In *The 32nd DAGM conference on Pattern recognition*, pages 202–211, Berlin, Heidelberg, 2010. Springer-Verlag.
- [97] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(9):1124–1137, September 2004.
- [98] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, November 2001.
- [99] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(2):147–159, February 2004.
- [100] K. McGuinness and N.E. OConnor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, February 2010.

- [101] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16(6):641–647, June 1994.
- [102] G. Friedland, K. Jantz, and R. Rojas. Siox: simple interactive object extraction in still images. In *7th IEEE International Symposium on Multimedia*, pages 253–260, December 2005.
- [103] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Transactions on Image Processing*, 9(4):561–576, April 2000.
- [104] T. Adamek. *Ph.D. thesis: Using contour information and segmentation for object registration, modeling and retrieval*. PhD thesis, School of Electronic Engineering, Dublin City University, 2006.
- [105] J. Weston, S. Bengio, and D. Grangier. Label embedding trees for large multi-class tasks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 163–171, 2010.
- [106] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531, June 2005.
- [107] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.
- [108] P. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang. Searching the world’s herbaria: A system for visual identification of plant species. In *European Conference on Computer Vision (ECCV)*, pages 116–129, 2008.
- [109] P. Bruneau, F. Picarougne, and M. Gelgon. Interactive unsupervised classification and visualization for browsing an image collection. *Pattern Recognition*, 43(2): 485–493, February 2010.
- [110] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1577–1584, 2011.
- [111] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. From region similarity to category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2665–2672, June 2011.

- [112] T. Dharani and IL. Aroquiaraj. A survey on content based image retrieval. In *International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME)*, pages 485–490, February 2013.
- [113] D. Zhang, F. Wang, Z. Shi, and C. Zhang. Interactive localized content based image retrieval with multiple-instance active learning. *Pattern Recognition*, 43(2): 478–484, February 2010.
- [114] B. Thomee and M.S. Lew. Relevance feedback in content-based image retrieval: promising directions. In *13th annual conference of the Advanced School for Computing and Imaging*, pages 450–456, Heijen, Netherlands, 2007.
- [115] C.B. Akgul, D.L. Rubin, S. Napel, C.F. Beaulieu, H. Greenspan, and B. Acar. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, 24(2):208–222, April 2011.
- [116] T. Havinga. Automatic severity assessment of hand eczema. Master’s thesis, University of Groningen, August 2010.
- [117] L. Ballerini, X. Li, R. Fisher, B. Aldridge, and J. Rees. Content-based image retrieval of skin lesions by evolutionary feature synthesis. In *Applications of Evolutionary Computation*, volume 6024 of *Lecture Notes in Computer Science*, pages 312–319. Springer Berlin Heidelberg, 2010.
- [118] L. Ballerini, X. Li, R. Fisher, and J. Rees. A query-by-example content-based image retrieval system of non-melanoma skin lesions. In B. Caputo, H. Muller, T. Syeda-Mahmood, J. Duncan, F. Wang, and J. Kalpathy-Cramer, editors, *Medical Content-Based Retrieval for Clinical Decision Support*, volume 5853 of *Lecture Notes in Computer Science*, pages 31–38. Springer Berlin Heidelberg, 2010.
- [119] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR ’03, pages 119–126, New York, NY, USA, 2003.
- [120] A. Yavlinsky, E. Schofield, and S. Ruger. Automated image annotation using global features and robust nonparametric density estimation. In *Image and Video Retrieval*, volume 3568 of *Lecture Notes in Computer Science*, pages 507–517, Berlin, 2005. Springer.
- [121] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *6th ACM international conference on Image and video retrieval*, CIVR ’07, pages 25–32, New York, NY, USA, 2007.

- [122] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562, The MIT Press, 55 Hayward Street Cambridge, MA 02142-1493 USA, April 2001. MIT Press.
- [123] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *12th annual ACM international conference on Multimedia, MULTIMEDIA '04*, pages 348–351, New York, NY, USA, 2004.
- [124] N. Zhou, W.K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(7):1281–1294, July 2011.
- [125] F. Yan, K. Mikolajczyk, J. Kittler, and A. Tahir. Combining multiple kernels by augmenting the kernel matrix. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, volume 5997 of *Lecture Notes in Computer Science*, pages 175–184, 2010.
- [126] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *European Conference on Computer Vision (ECCV)*, volume 7577 of *Lecture Notes in Computer Science*, pages 86–99, 2012.
- [127] X. Yu, T.C. Lik, Y. Yang, C. Fermuller, and Y. Aloimonos. Active scene recognition with vision and language. In *IEEE International Conference on Computer Vision (ICCV)*, pages 810–817, 2011.
- [128] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1681–1688, 2011.
- [129] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958, June 2009.
- [130] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *European conference on Computer vision (ECCV)*, pages 438–451, Berlin, Heidelberg, 2010. Springer-Verlag.
- [131] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2524–2531, 2011.

- [132] J. Ning, L. Zhang, D. Zhang, and C. Wu. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 43(2):445–456, February 2010.
- [133] O. Pauplin, P. Caleb-Solly, and J. Smith. User-centric image segmentation using an interactive parameter adaptation tool. *Pattern Recognition*, 43(2):519–529, February 2010.
- [134] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1832–1839, Barcelona, 2011.
- [135] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas. Automatic image annotation using group sparsity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3312–3319, June 2010.
- [136] I. Maglogiannis and C. Doukas. Overview of advanced computer vision systems for skin lesions characterization. In *IEEE Transactions on Information Technology in Biomedicine*, pages 721–733, September 2009.
- [137] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler. Automated melanoma recognition. *IEEE Transactions on Medical Imaging*, 20(3):233–239, March 2001.
- [138] E. Claridge, S. Cotton, P. Hall, and M. Moncrieff. From colour to tissue histology: Physics based interpretation of images of pigmented skin lesions. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 2488 of *Lecture Notes in Computer Science*, pages 730–738. Springer Berlin Heidelberg, 2002.
- [139] A. Safi, V. Castaneda, T. Lasser, and N. Navab. Skin lesions classification with optical spectroscopy. In H. Liao, P.J.E. Edwards, X. Pan, Y. Fan, and G. Yang, editors, *Medical Imaging and Augmented Reality*, volume 6326 of *Lecture Notes in Computer Science*, pages 411–418. Springer Berlin Heidelberg, 2010.
- [140] A. Green, N. Martin, J. Pfitzner, M. O’Rourke, and N. Knight. Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, 31(6):958–964, December 1994.
- [141] P. Schmid-Saugeona, J. Guillodb, and J. Thirana. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*, 27(1):65–78, January-February 2003.

- [142] P. Rubegni, G. Cevenini, M. Burrioni, R. Perotti, G. Dell'Eva, P. Sbano, C. Miracco, P. Luzi, P. Tosi, P. Barbini, and L. Andreassi. Automated diagnosis of pigmented skin lesions. *International Journal of Cancer*, 101(6):576–580, October 2002.
- [143] A.D. Hamilton and R.R.W. Brady. Medical professional involvement in smartphone apps in dermatology. *British Journal of Dermatology*, 167(1):220–221, July 2012.
- [144] C. Brodley, A. Kak, C. Shyu, J. Dy, L. Broderick, and A.M. Aisen. Content-based retrieval from medical image database: A synergy of human interaction, machine learning, and computer vision. In *The Sixteenth National Conference on Artificial Intelligence*, pages 760–767, 1999.
- [145] C. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A.M. Aisen, and L.S. Broderick. Assert: A physician-in-the-loop content-based retrieval system for hrcr image databases. *Computer Vision and Image Understanding*, 75(1-2):111–132, July 1999.
- [146] H. Muller, A. Rosset, J. Vallee, and A. Geissbuhler. Integrating content-based visual access methods into a medical case database. In *Medical Informatics Europe Conference (MIE)*, volume 95, pages 480–485, St. Malo, France, May 2003.
- [147] L. Savolainen, J. Kontinen, E. Alatalo, J. Rning, and A. Oikarinen. Comparison of actual psoriasis surface area and the psoriasis area and severity index by the human eye and machine vision methods in following the treatment of psoriasis. *Acta dermatovenereologica*, 78(6):466–467, November 1998.
- [148] E.F. Nakamura, A.A.F. Loureiro, and A.C. Frery. Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys (CSUR)*, 39(3):Article 9, September 2007.
- [149] X. Zhao, Q. Luo, and B. Han. Survey on robot multi-sensor information fusion technology. In *7th World Congress on Intelligent Control and Automation (WCICA)*, pages 5019–5023, June 2008.
- [150] A.A. Goshtasby and S. Nikolov. Guest editorial: Image fusion: Advances in the state of the art. *Information Fusion*, 8(2):114–118, April 2007.
- [151] I. Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone. Information fusion for computer security: State of the art and open issues. *Information Fusion*, 10(4):274–284, October 2009.

- [152] L.I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, Hoboken, NJ, USA, 2004.
- [153] T.G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2000.
- [154] G. Valentini and F. Masulli. Ensembles of learning machines. In M. Marinaro and R. Tagliaferri, editors, *Neural Nets*, volume 2486 of *Lecture Notes in Computer Science*, pages 3–20. Springer Berlin Heidelberg, 2002.
- [155] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
- [156] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.
- [157] O. Razeghi, G. Qiu, H. Williams, and K. Thomas. *Computer Aided Skin Lesion Diagnosis with Humans in the Loop*, pages 266–274. Machine Learning in Medical Imaging, 2012.
- [158] O. Razeghi, G. Qiu, H. Williams, and K. Thomas. Skin lesion image recognition with computer vision and human in the loop. In *Medical Image Understanding and Analysis (MIUA)*, Swansea, UK, pages 167–172, 2012.
- [159] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S.P. Khudanpur, D. Klakow, M.R. Krause, R. Manmatha, H.J. Nock, D. Petkova, B. Pytlik, and P. Virga. Joint visual-text modeling for automatic retrieval of multimedia documents. In *13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 21–30, New York, NY, USA, 2005.
- [160] O. Razeghi, H. Fu, and G. Qiu. Building skin condition recogniser using crowd-sourced high level knowledge. In *Medical Image Understanding and Analysis (MIUA)*, Birmingham, UK, pages 225–230, 2013.
- [161] O. Razeghi, Q. Zhang, and G. Qiu. Interactive skin condition recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2013.
- [162] J. Schofield, D. Grindlay, and H. Williams. *Skin Conditions in the UK: a Health Care Needs Assessment*. Centre of Evidence-Based Dermatology, University of Nottingham, Nottingham, UK, 2009.

- [163] R.J. Hay and L.C. Fuller. The assessment of dermatological needs in resource-poor regions. *International Journal of Dermatology*, 50(5):552–557, May 2011.
- [164] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *International Journal of Computer Vision (IJCV)*, 90(1):88–105, May 2010.
- [165] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 309–316, September 2009.
- [166] B. Scholkopf, A. Smola, and K. Muller. Kernel principal component analysis. In *Artificial Neural Networks (ICANN)*, pages 583–588. Springer, 1997.
- [167] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org>, 2008.
- [168] O. Razeghi and G. Qiu. 2309 skin conditions and crowd-sourced high-level knowledge dataset for building a computer aided diagnosis system. In *IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 61–64, April 2014.
- [169] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978, 2001.
- [170] R. Ashton and B. Leppard. *Differential diagnosis in dermatology*. Radcliffe Publishing Ltd, Abingdon, UK, 2005.
- [171] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145–175, May 2001.
- [172] C.C. Chang and C.J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, April 2011.
- [173] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In F.F. Soulie and J. Herault, editors, *Neurocomputing*, volume 68 of *NATO ASI Series*, pages 41–50. Springer Berlin Heidelberg, 1990.
- [174] J. Suckling et al. The mammographic image analysis society digital mammogram database. *Excerpta Medica. International Congress*, 1069:375–378, 1994.
- [175] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, November 1973.

- [176] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, volume 5099 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, 2008.
- [177] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698, November 1986.
- [178] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- [179] M. Torres and G. Qiu. Automatic habitat classification using image analysis and random forest. *Ecological Informatics*, 23(0):126–136, September 2014. Special Issue on Multimedia in Ecology and Environment.
- [180] P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, Cambridge, MA, 1974.
- [181] M.F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.
- [182] S. Kullback. *Information Theory and Statistics*. Wiley, New York, USA, 1959.
- [183] A. Mani, S. Napel, D.S. Paik, R.B. Jeffrey, J. Yee, E.W. Olcott, R. Prokesch, M. Davila, P. Schraedley-Desmond, and C.F. Beaulieu. Computed tomography colonography: Feasibility of computer-aided polyp detection in a first reader paradigm. *Journal of Computer Assisted Tomography*, 28(3):318–326, May-June 2004.
- [184] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, A.Y. Toledano, and K. Doi. Potential of computer-aided diagnosis to reduce variability in radiologists’ interpretations of mammograms depicting microcalcifications. *Radiology*, 220(3):787–794, September 2001.
- [185] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, USA, 2002.
- [186] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, September 1936.
- [187] E.C. Titchmarsh. *Introduction to the Theory of Fourier Integrals*. Clarendon Press, Oxford, UK, 1937.

- [188] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, January 1974.
- [189] C.K. Chui. *An Introduction to Wavelets*. Wavelet analysis and its applications. Academic Press, Boston, USA, 1992.
- [190] K.G. Beauchamp. *Applications of Walsh and related functions, with an introduction to sequence theory*. Microelectronics and signal processing. Academic Press, Orlando, USA, 1984.
- [191] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Knowledge Discovery and Data Mining*, pages 245–250, 2001.
- [192] S. Fidler, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(3):337–350, March 2006.
- [193] W. Zhao. Discriminant component analysis for face recognition. In *15th International Conference on Pattern Recognition (ICPR)*, volume 2, pages 818–821, 2000.
- [194] M. Yang, N. Abuja, and D. Kriegman. Face detection using mixtures of linear subspaces. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 70–76, 2000.
- [195] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.
- [196] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):711–720, July 1997.
- [197] B. Scholkopf, A. Smola, E. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.
- [198] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley, New York, USA, 1991.
- [199] W. Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5-6):823–837, September 1990.
- [200] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *The Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, December 2008.

-
- [201] M. Everingham, L. van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 results (voc2007), 2007.
 - [202] K. Bache and M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>.
 - [203] R.P. Gorman and T.J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89, 1988.